Modelling Risk Control Measures in Railways

Analysing how designers and operators organise safe rail traffic

Jaap van den Top

Modelling Risk Control Measures in Railways

Analysing how designers and operators organise safe rail traffic

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 19 april 2010 om 12:30 uur

door

Jaap VAN DEN TOP
civiel ingenieur
geboren te Arnhem.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. A.R. Hale
Copromotor: Dr. ir. H.M. Jagtman

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. A.R. Hale	Technische Universiteit Delft, promotor
Dr. ir. H.M. Jagtman	Technische Universiteit Delft, copromotor
Prof. Dr. –Ing. I.A. Hansen	Technische Universiteit Delft
Prof. Dr. –Ing. J. Pachl	Technische Universität Braunschweig
Prof. dr. ir. W.A.H. Thissen	Technische Universiteit Delft
Prof. J.R. Wilson	University of Nottingham
Dr. ir. A.A.M. Schaafsma	ProRail

Ir. T. Heijer heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

Dit proefschrift is het resultaat van onderzoek uitgevoerd door de vakgroep Veiligheidskunde van de TU Delft, faculteit Techniek, Bestuur en Management (www.tbm.tudelft.nl), en ProRail (www.prorail.nl), die daartoe hebben samengewerkt in het onderzoeksprogramma Next Generation Infrastructures (www.nginfra.nl). Het proefschrift is mede tot stand gekomen door ondersteuning van onderzoeksschool TRAIL (wwww.rstrail.nl).

ISBN 978-90-79787-15-9

Published and distributed by:
Next Generation Infrastructures Foundation
P.O. Box 5015
2600 GA Delft
The Netherlands
T: +31 15 278 2564
F: +31 15 278 2563
E: info@nginfra.nl
I: www.nginfra.nl

This research was funded by the Next Generation Infrastructures Foundation programme.

Keywords: railways, safety, process control, accident model

Cover template design by H. Slingerland BNO, Vlaardingen
Cover illustration: © 2010 Liset Terpstra
Copyright © 2010 by Jaap van den Top

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the author.

Printed in the Netherlands

"De exploitatie, de seinen, enz. alles moet als het ware met de ontwikkeling van het vervoer medegaan en steeds zorgvuldig worden overwogen, daar hetgeen sints jaren het bewijs van deugdelijkheid heeft gegeven, op een gegeven oogenblik kan blijken ten eenenmale onvoldoende te zijn en niet meer waarde dan een magtspreuk te bezitten."

"Operation, signals, etc. everything should closely follow the developments in transportation, and be constantly and meticulously re-examined, since that which has proved to be effective for many years may at a certain point show itself entirely inadequate and of no more value than a magic charm."

Verslag spoorwegongeval bij Warmond 14.8.1874, Raad van Toezicht op de spoorwegdiensten, 1875.
Report of the Warmond railway accident (14 Aug 1874), Railway inspectorate, 1875.

Preface and acknowledgement

Shortly before I graduated at Delft University, I was thinking about what kind of job I would like to take next. Two of my main interest areas were rail traffic management and rail safety, and I knew quite precisely that I wanted to combine them. However, no job appeared to be existing in which I could combine the two, until one day my thesis supervisor Tom Heijer came to me and told that he was joining forces with ProRail to start a project that would quite exactly deal with what I had in mind. Three preconceptions must be taken away right from the start, since doing scientific research is all about taking away preconceptions and poorly underpinned myths and beliefs, and about replacing these by observable data, explaining mechanisms and valid conclusions.

The first misconception was taken away even before this research actually started. Doing a PhD is not the boring, lonely, poorly paid, far too specialistic and dusty job without practical implications, as many people often appear to think. So I happily accepted the offer and an intensive quest for understanding phenomena that were obscured until that moment could get started. Why does the meticulously planned train timetable still lead to delays and operational problems? And how can railway accidents still happen, if technical systems and rules are made such that a single human error normally has no consequences? There seemed to be a loophole which cried out to be discovered. Now that you are reading this, the results of that quest are lying right in front of you. In order to enable a broad diversity of readers to understand this book, jargon, abbreviations and mathematical formulas are used only where this is absolutely necessary. Citations from originally non-English documents were translated by me, unless otherwise indicated. In some cases however, it was not possible to find a translation with the very same subtilities that can sometimes implicitly carry a lot of meaning. As such, the language in which a research is performed partly influences its outcomes.

The idea that research has no practical implications must certainly be gainsaid, since the knowledge I gained during this research was also asked for by society: on their request, it took me to the court and to the parliament to give explanations. Some readers of this thesis will not be happy with the findings of this thesis, since they will fear that their organisations' image will be damaged. If you are one of them, please take into consideration that making errors and admitting them is a first necessary step to identify what goes wrong and to exploring options

for how to improve. That is exactly what this thesis does. It is much worse not to admit errors, not to learn from them and to repeat them again and again.

Secondly, many people think that doing a PhD is a one-man job. This, too, was not the case. Thank you Andrew, Ellen, and Tom, for your repetitive and constructive comments and discussions. I am still impressed about your ability to re-read texts again, again, again and again and still being able to motivate me. Thanks go also to all of the Safety Science group colleagues for your very personal contribution to making my work a pleasant endeavour. My regards also go to ProRail, who provided me not only with budget and information to carry out this research, but also with an almost incredible number of supervisors: Toine Houben, Robin Hemelrijk, Hans Kuijlen, Philip de Graaf, Matthijs Jager, Maarten de Klerk, Frank Fiselier, Linda Wright, Gert Dijkstra, Laurens Berger, Maaike van Mourik, John van Engelen and Marcel Hendriks Boers, thanks for taking the effort to guide me through your large and complex organisation. Five ProRail employees, Alfons Schaafsma, Bruno van Touw, Egbert van Goor, Erik van der Meer, and Rolf Hetterschijt, were so kind as to review the concept of this thesis. Hans Klarenbeek, you were the first one to offer such a review and, to my dismay, you could not live up to your promise. I still have great memories of our discussions and your willingness to share your amazing amount of knowledge.

A most special word of thanks goes to all those kind railway staff members who were so accessible and enthusiastic to provide me with information in long interviews and to speak openly about their work. Their contribution to this work cannot remain unmentioned. However, since I promised not to disclose their names they must remain anonymous. The people I mean here know who they are and I want to thank them very, very much; without your information this thesis would not have been possible at all. Other, more informal, but not less useful discussions I had with Arco Sierts, Bauke Steenhuisen, Jan Lemmens, Jelle van Luipen, John Kingston, Marco Lüthi and so many others. Thus, it can hardly be stated that doing a PhD comes down to one person.

Also in the private sphere, there were a large number of people who had the strong belief that I could manage this. They were there when I needed them, either to give support, to give some useful distraction, or to share joyful moments. Karin, Wim, Vincent, Michiel, Wubbo, all friends from M-Jong, Marije, Margo and Jan and especially Shimriet: you were a great help.

The third and last preconception that turned out to be false is that every PhD candidate sooner or later gets confronted with a writer's block. As you may see from the number of pages, I liked writing the book, and I hope you will enjoy reading it just as much – especially if you happen to be one of those people whose responsibility it is to do something useful with the results.

Jaap van den Top
Delft, February 2010

Contents

PREFACE AND ACKNOWLEDGEMENT I

PART I: BACKGROUND KNOWLEDGE

1 INTRODUCTION 3
1.1 Short overview of this thesis 3
1.2 SPAD: a striking, unintended and possibly dangerous event 3
1.3 Societal developments: better and more services 5
1.4 Increased network load: less reliability and more SPADs 7
1.5 Research objective and approach 9

2 BASICS OF RAILWAY ORGANISATION AND OPERATION 15
2.1 Introduction 15
2.2 Players and markets in the railway sector 15
2.3 Control processes in railway operation 26
2.4 Conclusion 39

3 ANALYSIS OF EXISTING RISK CONTROL MODELS 41
3.1 Introduction 41
3.2 Basics of safety science 41
3.3 Overview of existing approaches to explain accidents 44
3.4 Conclusion 50

PART II: THEORETICAL FRAMEWORK

4 AN INTEGRATED RISK CONTROL MODEL: THE SAFE ENVELOPE OF OPERATIONS 55
4.1 Introduction 55
4.2 Development of the safe envelope so far 55
4.3 Terms and definitions 57
4.4 Balancing attention between safety and quality 61
4.5 Further development of the safe envelope in a railway context 61
4.6 Illustration of an accident in the safe envelope model 66
4.7 Lessons learned from applying the safe envelope model 69
4.8 Discussion 73
4.9 Conclusion 74

5 PROCESS CONTROL WITHIN THE SAFE ENVELOPE 77
5.1 Introduction 77
5.2 Cybernetics 77
5.3 Detect – Diagnose – Act: three general control subprocesses 81
5.4 Same subprocesses in other fields 83
5.5 Discussion 86
5.6 Conclusion 89

6 THE THEORETICAL FRAMEWORK: A MIX OF MODELS AND METHODS 91
6.1 Introduction 91
6.2 Combining the models 92
6.3 Collecting and processing of data using a grounded theory approach 93
6.4 Conclusion 99

PART III: UNDERSTANDING THE CURRENT SITUATION

7 THE TRAIN CONTROL PROCESS 103
7.1 Introduction 103
7.2 Signalling from a control perspective 104
7.3 The Dutch signalling system and its practical use 114
7.4 Driver human factors 124
7.5 Train control in safe envelope terms 138
7.6 Conclusion 140

8 THE TRAFFIC CONTROL PROCESS 143
8.1 Introduction 143
8.2 Route setting: assigning infrastructure to users 144
8.3 Rescheduling infrastructure use 151
8.4 Traffic control human factors 157
8.5 Ambiguous practices: cooperation with other actors 159
8.6 Traffic control in safe envelope terms 163
8.7 Conclusion 165

9 INFLUENCES FROM DESIGN 169
9.1 Introduction 169
9.2 Designing the service offer 170
9.3 Designing and supporting the operational processes 173
9.4 Design of safety systems 179
9.5 Design and management in safe envelope terms 206
9.6 Underlying causes: management responsibilities 208
9.7 Conclusion 215

PART IV: DISCUSSION AND FUTURE

10 HOW TO IMPROVE SAFETY AND QUALITY OF SERVICE? 221
10.1 Introduction 221
10.2 Short summary of current operational control problems 222
10.3 Reflecting on the theoretical framework 223
10.4 Unravelling the problem (A): Encourage operators to report surprises 227
10.5 Unravelling the problem (B): Improve cooperation between operators 228
10.6 Unravelling the problem (C1): Learning from actual operations 232
10.7 Unravelling the problem (C2): Traffic control improvements 237
10.8 Unravelling the problem (C3): Train control improvements 239
10.9 Discussion 243
10.10 Conclusion 249

11 CONCLUSIONS AND RECOMMENDATIONS 251
11.1 State of the art in research and need for additional knowledge 251
11.2 Framework to understand process control, safety and quality of service 252
11.3 Explanations for the current quality and safety problems in railways 254
11.4 Improving risk control measures and their modelling in railways 256
11.5 Recommendations 258

REFERENCES 261

PART V: APPENDICES

APPENDIX A EXAMPLES OF CURRENT PROJECTS 279
APPENDIX B INTERVIEW QUESTIONNAIRE 287
APPENDIX C SUBSEQUENT MEANINGS OF THE YELLOW SIGNAL ASPECT 289
APPENDIX D LEARNING FROM SPAD DATABASES 291
APPENDIX E COSTS AND BENEFITS 297
GLOSSARY 303
SUMMARY 305
SAMENVATTING 311
CURRICULUM VITAE 319
NGINFRA PHD THESIS SERIES ON INFRASTRUCTURES 321

PART I: Background knowledge

1 Introduction

1.1 Short overview of this thesis

In brief, this thesis addresses the question as to how the controllability of the railway system can be improved in order to increase both the level of safety (especially, as an example, Signals Passed At Danger, or SPAD) and the quality of service. In order to improve, it is necessary to understand which mechanisms underlie quality and safety problems. This is especially necessary because the current governmental policy implies that more trains will have to be handled on the existing rail network. This thesis therefore aims to understand how the railways are controlled and to provide suggestions as to how the controllability can be improved, so that the SPAD risk is reduced and quality of service can increase. This chapter serves to describe developments in society and in railways, to introduce the research question and to define the scope of the thesis.

1.2 SPAD: a striking, unintended and possibly dangerous event

1.2.1 Why train drivers do not want to pass a red light

A red light in railways means that an approaching train is not allowed to pass the light and should be brought to a stop in front of it. In contrast with road traffic, where passing a red light may result in shorter travel times as compensation for the risk taken, this is not beneficial at all for a train driver. The following reasons are the conclusion of an expert meeting organised by the Dutch railways and illustrate why it is not beneficial for drivers to run through a red light (Nederlandse Spoorwegen 2 March 2007):

- the train which was involved in a SPAD will be stopped by traffic control, and a number of operating procedures have to be completed before the train can proceed. A SPAD thus causes delay;
- the driver who was involved in a SPAD is known. He will be subject to a thorough re-examination on various psychological aspects;
- a SPAD is felt to be a serious proof of the absence of professional skills by train drivers;
- a driver risks a train accident and thus his own life, and those of the passengers in the trains involved.

1.2.2 Dutch SPAD history

So, one can safely assume that a train driver wants to stop in front of a signal at danger. Nevertheless, the figure of reported SPADs has almost doubled from just under 150 yearly during the period 1981 – 1995 to an average annual number of approximately 270 SPADs in 2004-2008, as Figure 1.1 shows (Railned Spoorwegveiligheid 2002; Inspectie Verkeer en Waterstaat 2009c). The number of accidents following a SPAD had a similar development from 1.3 per year on average during 1981-1995 to 3.4 per year in 2004-2008 (see Figure 1.2).

1.2.3 Difficulties in the explanation of SPADs

The increase cannot be explained by an increase in network usage. The amount of train kilometres rose from 112 million in 1981 to 136 million in 2006, an increase of 21% (Thompson 2001; ProRail 2007). One might assume that every train added to the network can encounter red signals itself, but it can also cause red signals for other trains, so that the SPAD rate would increase roughly quadratically; but even then, only a 47% growth could be explained. Underreporting in the period before 1995 can also not be a full explanation, since the number of accidents also increased, and it is unlikely that these too have been under-reported. Apart from that, an increase in network usage would, even in the best case only be able to explain the *increase* in SPAD numbers, but not the phenomenon itself: a causal explanation for SPADs is required, since trains do not run through red signals just because there are so many of them.

In a small number of cases, the driver is not actively involved in SPADs. Approximately 13% of SPADs can be attributed to technical problems, mostly 'slippery tracks', and 7% of SPADs occur when rail vehicles roll through a red signal when they were wrongly parked (Inspectie Verkeer en Waterstaat 2008h). This means that in the vast majority of cases (approximately 200 every year) the driver, who would be willing to stop and who also had the means to do so,

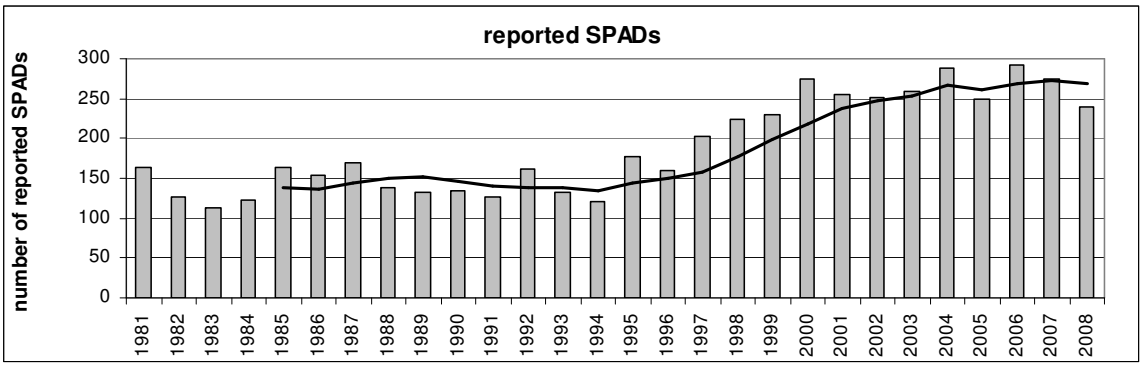


Figure 1.1: Number of reported SPADs per year, including 5 year moving average.

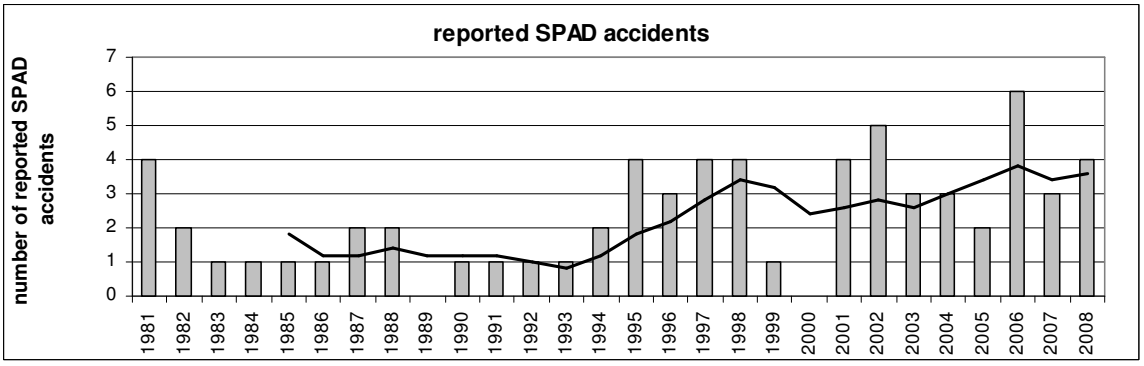


Figure 1.2: Number of reported SPAD accidents per year, including 5 year moving average.

runs through a red signal. Indeed, most recent SPAD accidents in the Netherlands did not happen during heavily disturbed operations, but during normal operations with small delays.

1.3 Societal developments: better and more services

1.3.1 Political wishes: quality, reliability, safety

Several resolutions in the Dutch parliament ask for better performance of the Dutch railway system. They ask the government for example to:

- improve intercity services, so that they will become faster, safer and more reliable (Van Gijzel et al. 2000);
- to see whether newly available technology can be cost-effectively used to improve the railway network, so that more and faster services can be offered and more passengers can be transported (Slob et al. 2000)
- make train services better, quicker, more comfortable and reliable (Dijksma and Van Hijum 2006);

In summary, when the parliament speaks of 'better utilisation of the railway system' they mean that the railway system should provide better service levels to customers.

1.3.2 Governmental strategy: better utilisation of the existing rail network

In its current transport plan 'Nota Mobiliteit' (White paper: 'Mobility'), the Dutch Ministry of Transport puts emphasis on two issues:

- the door-to-door reliability of all transportation modes should be improved; and
- due to the high investment costs of new infrastructure, existing infrastructure should be used in a smarter way. Only if that is not possible, should new infrastructure be built.

A previous objective was to change the modal split towards societally more preferable transport modes (i.e public transport), however this goal has now been explicitly abandoned and the government now wants to exploit each transport mode's specific advantages: cars for criss-cross trips and public transport for mass mobility. Similarly, the government does not want to limit their citizens' mobility, since mobility is considered an essential need and associated with economic growth (Ministerie van Verkeer en Waterstaat 2004a).

Growth rates for public transport use have varied considerably in different forecasts and the real growth has been even stronger:

- in response to the (rejected) predecessor of the Nota Mobiliteit, the companies in the Dutch rail sector and the Ministry of Transport jointly published a plan that they called 'Benutten en Bouwen' (Utilise and Build). This plan assumes that the rail transport demand in the Western, busiest part of the Netherlands would be doubled by 2020 from the figure in 2003; in the other parts, growth of 40-70% should be strived for, which corresponds to growth rates of 2% – 4% per year (Project Benutten & Bouwen 2003a);
- only one year later, the forecasted growth rate in the Nota Mobiliteit was less ambitious and set to approximately 1% per year;
- since this 2004 decision, rail transport in the Netherlands has been growing with 5% per year and the government is now willing to retain this grow rate (Ministerie van Verkeer en Waterstaat 2007).

It will be clear that growth is expected and that this growth must be handled on the existing rail network as far as possible. With the 'programme for high-frequency rail transport' (programma hoogfrequent spoorvervoer), an updated and more concrete plan to increase train frequencies has been published in 2008 (ProRail et al. 2008). It is proposed to increase train frequencies to 6 intercity services and 6 local services per hour on fixed corridors; also several freight trains per hour must be able to run. Compared to the current situation, this is

approximately a 50% increase in train frequencies. As section 1.4 will show, such increased traffic volumes give further need to investigate the SPAD incidents. A fundamental question is whether the current control philosophy will be able to cope with such intense traffic.

1.3.3 Avoiding unproductive minutes of the production system

The idea behind better utilisation is to avoid unproductive minutes of the means of production. Maarten van Eeghen, deputy secretary of the Ministry of Transport, clarified this in an interview for a public relations magazine of Nederlandse Spoorwegen when the new future plan 'Benutten en Bouwen' was introduced: '*Look at other industries, such as the maritime or aviation industry. All logistic processes are aimed at keeping those ships and planes moving. Every minute that they are at a quay or on the ground costs money. For trains and the expensive rail infrastructure, this is not different. Use it, use it, use it. That's what it is all about'* (Nederlandse Spoorwegen november 2003). Indeed, research showed that self-financing public rail transport can be reached if 20,000 passengers per day use a line, and that short running times are one of four main factors to explain cost-effectiveness (Hofstra 2005). Short running times mean that a more attractive service is provided, hence higher incomes can be achieved, but they also lead to short circulation times for rolling stock and personnel, and hence to lower production costs.

However, when 'Benutten en Bouwen' was 'marketed', it was not a reduction of travel times or avoidance of unproductive time that became central, but the idea that no infrastructure should be built if the existing infrastructure could handle a higher number of trains. In the urge to increase network capacity, the idea of 'squeezing' as many trains as possible over the existing infrastructure took over from its original intention. Since the maximum number of trains on a line is reduced by the fact that fast trains cannot overtake slow trains on double track lines, several proposals were made to change the timetable such that more trains could be handled without changes to the infrastructure, starting seriously with a study by Railforum (Railforum Nederland 2001). It was proposed to 'homogenise' rail traffic, which basically came down to making intercity services stop more often so that the running time differences between fast and slow trains would be reduced. This however leads to more unproductive minutes, since nobody really asked for slower intercity services. Instead, the parliament asked, even repeatedly, for quicker services. In summary, 'better utilisation' has gradually shifted from realising better services to the question how additional trains can be squeezed onto the existing rail network.

Already in 2002, economic research institute Nyfer has warned for problems of intensified network use, since the Dutch railway network is already one of the most intensively used ones. When production is intensified, the average costs may then rise rather than fall due to congestion and waiting times. Nyfer was therefore sceptical about an even higher network utilisation, although the institute acknowledged that it may become possible to realise a higher track utilisation without an increase in costs with technological developments or smarter ways of planning (Poort 2002). Such options, albeit hardly ever mentioned, would have been to introduce more passing tracks or to homogenise traffic by speeding up slower trains (Onderwater and Leusden 1999; Hansen 2003). A study into best practices for capacity enhancement by Booz Allen Hamilton (2007) indicated that harmonisation of speeds leads to sub-optimisation, in order to create the required capacity. This, so they say, has a limited viability.

This thesis aims to find out whether and how improved process control can be used as a lubricant to smoothen the operational processes.

1.3.4 Capacity: number of trains or customer demand?

What goes wrong in the abovementioned situation is that the capacity of the railway is (implicitly) referred to as the number of trains per hour that the infrastructure can handle. However, as will be further explained in section 2.2.3, such train paths are a half-product and not the end-product to be achieved, which is meeting customer's travel needs or demands. Train paths or 'capacity' should not be transformed such that they no longer fit the market demand: *"Line capacity measured in 'trains per hour' is really a technical abstraction, useful for comparing some details of signalling schemes, but for little else"* (Barter 2008).

Why does the word capacity lead to difficulties? A train path is *'the infrastructure capacity needed to run a train between two places over a given time-period'* (Directive 2001/14/EC). However, '*a unique, true definition of capacity is impossible'* because the capacity of the infrastructure is dependent on the types of train paths that the train operating companies ask for (UIC 2004). Since it depends on what is being asked for capacity can therefore only be qualified as being adequate if no one is complaining. This was already mentioned by Schaafsma and Weits (1996), and this view is now also made part of formal capacity allocation regulations: *'Where (...) it is not possible to satisfy requests for infrastructure capacity adequately then the infrastructure manager must immediately declare that element of infrastructure on which this has occurred to be congested'* (Directive 2001/14/EC). Capacity does not at all relate to the number of trains that can be squeezed over a line, but about the ability of a (rail) system to satisfy the customer's demands. This stays rather close to the original meaning of the word 'capacity', which in fact indicates whether something is capable or suitable to perform a certain task. This is also reflected in the UK Institution of Railway Operators' definition of network capacity: *'[Capacity is] the number of trains that can be incorporated into a timetable that is conflict-free, commercially attractive, compliant with regulatory requirements, and can be operated in the face of anticipated levels of primary delay whilst meeting agreed performance targets'* (Department for Transport 2007). Thus, many more criteria are relevant to define the capacity than just the number of trains that can be scheduled: if quality of service or traffic safety is insufficient, this also indicates that the network capacity – in a broader sense, e.g. including those who define how the capacity of the means of production should be used – is insufficient. Given the difficulties and the confusion that comes with the word capacity, it is suggested not to use the word and always to refer to train paths, quality of service or safety instead, since that is what the term 'capacity' generally implicitly indicates. Capacity can be used, however, as an abstract term for unassigned means of production that can still be put into service once this is required.

1.4 Increased network load: less reliability and more SPADs

1.4.1 Expected problems of increased loads

Another fundamental issue in this respect is how 'better utilisation' can be measured, since this requires in the first place a clear definition of 'good utilisation'. Under pressure of better (read: higher) track utilisation, one might be tempted to cut out spare capacity, i.e. to reduce the number of spare train paths. However, there will always be (minor) variations in arrival and running times of trains. Also, late trains cannot use their original path and must claim a new one. It is necessary to leave some spare capacity in the timetable in order to handle these common disruptions. Using all spare capacity may put quality of service at risk. Possibly, also safety can be influenced when trains have to make many unplanned stops because no path could be provided in time. As a paradox, one may ask whether track utilisation is indeed 'better' if quality and safety are likely to decrease. There is a continuous conflict between (utilisation of) spare capacity on the one hand and quality of service and safety on the other.

Since this conflict is caused by the possibility of disturbed operations as opposed to the sure wish to utilise the system as much as possible, easing this conflict is possible by enhanced capability to handle disturbances.

The higher the network utilisation, the smaller the disturbances that will already affect other trains and, via this interaction, both the quality of service and the safety level are at risk. It can be expected that trains will have to wait for each other more often, as the probability is increased that a necessary track section is still being occupied by another train (Schwanhäußer 1974; Huisman et al. 2002; Poort 2002; Hansen and Pachl 2008). Thus, signals will show a 'stop' aspect more often. This will not only lead to waiting times (either planned or unplanned), but it will also create more 'opportunities' to produce a SPAD.

This consequence of a higher network utilisation appears to be conflicting with the demand for a higher reliability. Also, it appears to contradict the policy on rail safety, which states that the safety level should remain at least on the current level: a maximum of 0,15 fatalities per 10^9 passenger kilometres, excluding suicides (Ministerie van Verkeer en Waterstaat 2004b). Since disturbances can never be fully taken away, it is necessary to learn to cope with them and manage the interaction between trains in a smarter way.

1.4.2 Current solutions

In the rail sector, quality problems are usually solved by altering the timetable, but this cannot be a solution for all disturbances that occur during operations, as the exact appearance of the disturbance can only be known minutes in advance (Schaafsma 2001; Van den Top 2005; Projectteam Samensporen 2006). For safety problems a technical solution is often seen as the ultimate answer, as several examples show:

- ProRail's traffic control department stated: *"not applying such measures [to prevent SPADs] by the traffic controller will often be underpinned with the 'punctuality' argument (…). To prevent such undesired side-effects and (new) risks, measures should by definition be taken in the safety systems (…). In doing so, exclusion of human error is always the leading principle when safety measures are formulated. Principally, all measures must be anchored in technology. Only if it can be demonstrated that this is impossible, can additional regulations be brought under discussion"* (Sanders 2005).
- ProRail's design rule AV101 for railway safety gives the following definition for the protection of rail traffic: *'With protection of rail traffic is meant the promotion, as much as possible, of safe and efficient devices, possibly completed by regulations'*[1] (ProRail 2005a). In other words, not safety itself is defined but the way in which it should be reached, which is through as much technology as (economically) possible.
- The railway inspectorate stated in reaction to a SPAD accident that: *'this violation is caused by human failure and will never be completely controllable (…). The inspectorate assumes, that the number of signals passed at danger can only be greatly reduced by technical solutions'* (Inspectie Verkeer en Waterstaat 2007b). This statement was made even though the inspectorate concluded that the main cause of the SPAD was a misplaced signal and the inspectorate proposed to add ergonomic requirements to signal placement norms.

[1] During the review process of this thesis ProRail discovered that the original meaning of AV101, before it was retyped in order to get an electronic version, read *'with protection of rail traffic is meant the promotion, as much as possible, of safe and efficient rail traffic processes by means of devices, possibly completed by regulations*' (emphasis added by author). This has now been corrected and will be issued in the next version of the design rules.

Apparently, in railways there is a general belief that humans can fail and that technology performs better. However, these technical safety systems are expensive (Klerk 9 July 2004; Dijkstra 2004; Ministerie van Verkeer en Waterstaat 2004b). Moreover, the strong focus on technical safety systems has resulted in not much effort being spent on human factors, although this has received more attention in the last decade (Wilson and Norris 2005), so that knowledge about why humans actually fail and under which circumstance is only relatively recent. Klerk and In't Veld (2005), respectively CEO and board member of ProRail, also indicate that the railways tend to look for technical solutions, but that there must as well be other possibilities to improve safety performance.

1.4.3 New solutions require broader understanding

The aim of the preceding is not to say that technology itself is necessarily bad, but that the one-sided focus on technology risks that other developments get out of focus. Maybe, technology should not be regarded as a replacement for failing human operators, but they should be seen as two complementary parts of one railway system. In other industries, there has been an increasing role for real time control when infrastructures were unbundled (de Bruijne 2006; van Eeten et al. 2006). Also in railways, it is the task of operational staff to keep the system going, so that transport services can be delivered in a safe way even when unforeseen disturbances occur. Traffic controllers, in co-operation with train drivers, know what is possible in a certain situation and what is not; they surely want the train to stop at signals showing red. Different studies in railways show that improvements in safety or performance can be achieved when traffic controllers (Kauppi et al. 2006; Wilson et al. 2007a) and train drivers have better support in their work (Kecklund et al. 2001; Wilson et al. 2007a). In aviation, traditionally more emphasis is put on human behaviour and as a result, knowledge exists about how systems should be designed in order to improve human performance (Endsley and Kiris 1995; Endsley 1999). Finally, a recent insight in safety science comes from Leveson (2004), who states that accidents are a result of a failing control system by which she also includes human elements.

1.5 Research objective and approach

1.5.1 Research question

The preceding sections give sufficient support to the belief that improvements in process control are likely to improve quality of service and safety of railways. However, the influence of process control on these aspects is not yet fully understood. As the current 'default' solutions (timetable changes and safety systems) appear not to bring a full solution, it is necessary to investigate how safety and quality of service can be effectively managed by improving process control. This leads to the following research question:

In what way could real time process control be improved to positively influence safety and quality of service in rail traffic operation?

1.5.2 Societal relevance

To do this, this research aims to find causal explanations for the mechanisms that lead to quality and safety problems. If such insight can be gained, measures to improve the situation can be taken more specifically. It will appear throughout this thesis that several improvements in process control are possible and that a higher network usage, the quality of service requirements and the safety requirements do not necessarily conflict. However, this requires a deeper and broader understanding by those who design and operate the railway system.

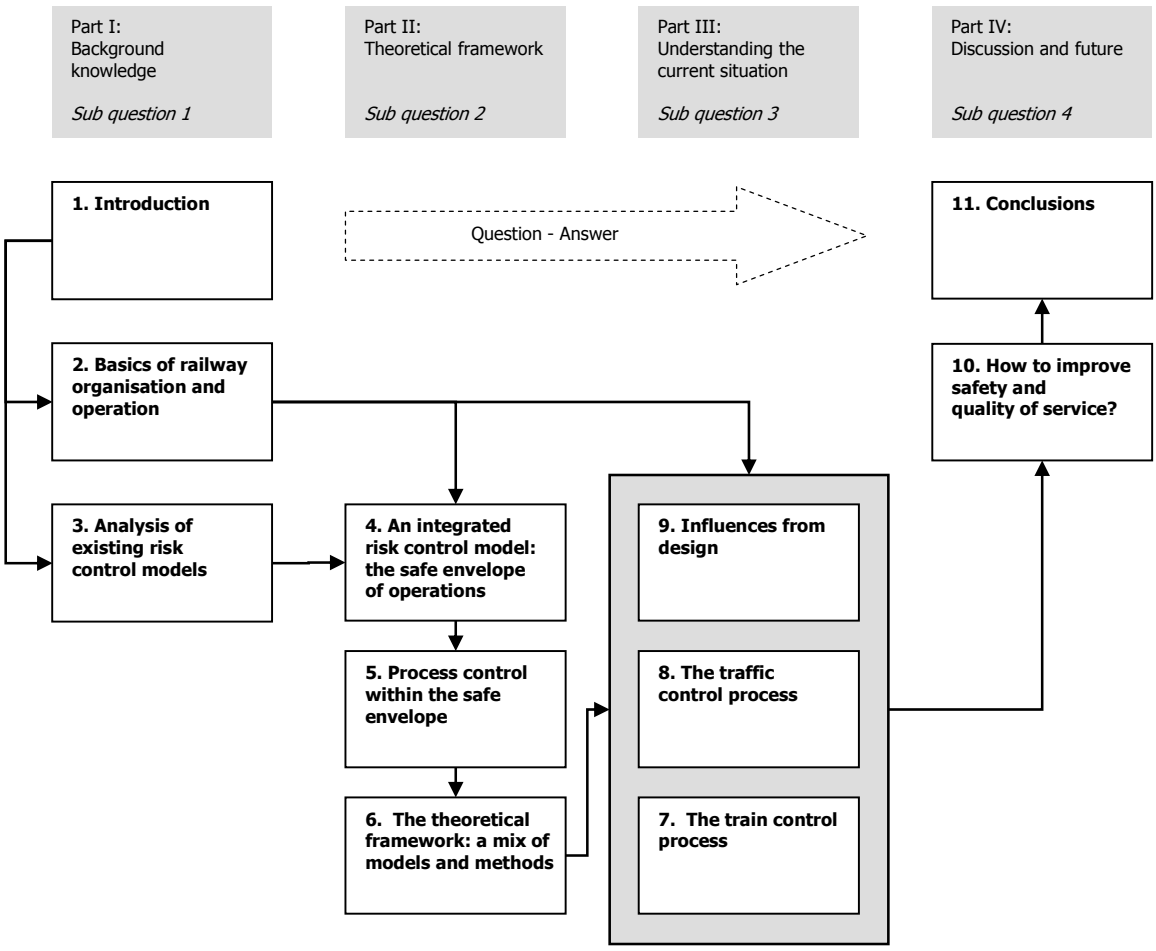


Figure 1.3: Overview of this thesis.

Problems during operations must be freely reported and solved at the appropriate level: in design, operations or a combination of both.

1.5.3 Research approach and sub questions

This section will describe the approach which is taken to answer the research question in more detail. In doing so, it will also describe the structure of this thesis, which is inherently linked to the research method (see Figure 1.3). This thesis is divided into four parts. Each part covers one of the sub questions that will be stated below.

Firstly, it is necessary to understand the basic concepts of railway operation and safety science and – since no clear answer to the research question is easily available – to see which gaps in knowledge exist. The first sub question therefore is:

1. **What is the state of the art in railway operation research and safety science and what additional knowledge is required to answer the research question?** (Chapters 1 to 3)

Since the main question aims to relate process control to railway safety and to quality of service, these three concepts must be linked in one theoretical framework. However, as Chapter 3 will show, current accident models hardly ever take production goals and safety into account simultaneously, nor are they fully suitable to model the influence of process control on a dynamic system, such as railways. Therefore, a new framework had to be developed. This leads to the second sub question:

2. **Which framework is needed to understand how process control influences safety and quality of service?** (Chapters 4 to 6)

To overcome the limitations of current accident models, Hale et.al. (2007) proposed a rough outline of a new model, called the safe envelope of operations, which however required further development. This development took place during the work done in this thesis. The terms and definitions on which the model is based were detailed in a more systematic way. Also, the model was extended in order to incorporate not only the safety boundaries that the operator should try to avoid, but also the system goals that he should try to achieve. An elaborate example in which the framework is illustrated, before being applied in a later stadium of this thesis, was also worked out. The results of this can be found in Chapter 4.

Because the safe envelope puts much emphasis on the controllability of the system, this topic was further explored, from technical, human factors and organisational perspectives (Chapter 5). This theoretical part combines accident models, control theory, cybernetics, human factors and management science. The development of the model building and also the subsequent study of controllability were also informed by extensive interaction with a range of projects and discussion with the people within the rail system dealing with them. Many of these projects are used as examples either in the text or in appendices.

As a next step, it is necessary to identify to what extent the current railway system fulfils the criteria of the framework. Analysis of scientific literature and professional documents took place and a total of 30 interviews took place with drivers, traffic controllers and several designers, as will be described in Chapter 6. Wherever necessary, design changes or new designs were also analysed. These data served a dual purpose: in the first place, these were analysed using a grounded theory approach (see Chapter 6) to derive underlying causes for performance problems from the data in an as much as possible unbiased way. Secondly, the data were analysed using the safe envelope framework. Both analyses served to be able to answer the third sub question:

3. **Using this framework, is it possible to find process control related explanations for the current quality and safety problems in railways?** (Chapters 7 to 9)

Once explanations have been found and a better understanding of SPAD and quality problems is available, the next step is to formulate a range of improvements aimed at making changes where the current situation does not match the requirements from the theoretical framework, improvements. Thus, the last sub question is:

4. **Given this framework, how can risk control measures and their modelling in railways be improved?** (Chapters 10 and 11)

Both safety and quality can be seen as outcome of a control process and occurring problems as failures of this. Improvements in the control system (which also comprises management) such as closing the feedback within operations, but also from operations to design and management, will be necessary to solve the problems.

1.5.4 Definitions

In order to avoid confusion, it is important to clarify what is meant by different terms. The following definitions will be used in this thesis.

- **Process control**: the process of causing a system variable to conform to some desired value, called a reference value (Franklin et al. 2006).

- **Real time**: A process is controlled in real time if the reaction time in response to an event is shorter than the maximum time imposed by the process that must be controlled. Real time control is necessary to avoid that the control system lags behind the process it controls. It therefore serves to avoid instability (Encyclopædia Britannica Online 01 Sep. 2009). Example: in an antilock braking system, the braking force must be reduced before the wheel blocks.
- **Real time process control:** in this thesis mainly used to indicate the work of the traffic control centres and train drivers.
- **Tolerance**: the allowable deviation from the desired reference value.
- **Plan(ning)**: A plan is a number of coherent measures in order to achieve certain goals. Planning is a (methodical) tool for control. Planning can be described as taking a coherent series of decisions such that the course of events will be influenced (de Leeuw 2001).
- **Safety**: freedom from those conditions that can cause death, injury, occupational illness, or damage to or loss of equipment or property (MIL-STD-882D 2000).
- **Reliability**: the ability of an item to perform a required function, under given environmental and operational conditions and for a stated period of time (ISO8402 1986).
- **Quality**: the degree to which a set of inherent characteristics fulfils requirements (ISO9000 2000).
- **Service**: non-material equivalent of physical goods. A product or a service is a (temporary or permanent) change in the customer and/or the context, which fulfils a need. A service is intangible, cannot be stored, often there is coproduction (the customer coproduces, otherwise no production is possible). A service can only be assessed after it has been delivered and the controllability of the result is only limited (de Leeuw 2001).
- **Train path**: a planned trajectory in time and space in which infrastructure is reserved for a specific train movement. A path is a service delivered by an Infrastructure Manager to a Train Operating Company.
- **Route**[2]: physically reserved infrastructure sections for a specific train (thus being the operational equivalent of a train path). Note that a *path* has dimensions in time and space; a *route* has only a spatial dimension; it can only exist as of the moment that it has been set.

1.5.5 Scope, limitations, assumptions

The scope of this study is on the safe and efficient provision of routes for trains. These are provided by traffic controllers (employees of the Infrastructure Manager) and used by train drivers (employed by a Train Operating Company).

The main safety hazard is a train collision, for which a SPAD is a precursor. Many railways have automatic train protection systems that stop trains when they are running through a red signal. Although not all of these systems can totally prevent that the train is running through a red light, they usually do warn the driver or apply the brakes. However, the Dutch system ATB (Automatische TreinBeïnvloeding) does not provide such functionality, since it only makes sure that a driver starts to brake when approaching a red signal, but the system does not check whether this is sufficient to stop in front of the red signal.

The main quality problem is that of insufficient reliability of timely train path provision. Since these problems show themselves first during operations, this will be the prime focus of this study. Investigating the influence of organisation, management and equipment design and provision may be necessary and will become clear when a problem or its possible solution point in that direction.

[2] In Dutch, this word is translated with “rijweg”; this may be confusing since the term route is in some cases used in the traffic control system (VPT).

This study will not investigate what acceptable risk and quality levels are. This is largely a matter of politics and market forces. Since this study's focus is on prevention of accidents in the railway system, it does not deal with mitigation or recovery of the system after an accident has happened. For the same reason, external safety (the risk imposed by the railway to its surroundings) is outside the scope of this study. Also vandalism and terrorism will not be considered.

There are safety issues that should be dealt with if train frequencies are increased, but that are not addressed in this research, basically because they cannot be influenced by the traffic control centre and/or the train drivers. These are listed below, with a reason for exclusion:
- Track maintenance (an extremely relevant issue, which should be able to be dealt with by extending the current approach, but which could not be tackled in the time available);
- Rolling stock maintenance (this hardly has any relationships with real time traffic control);
- Safety of level crossings (this is more a matter of road users and level crossing design, and can hardly be influenced by traffic controllers or train drivers);
- Safety of passengers on platforms (stumbling etc.; safety measures do not depend on real time traffic control[3]).

Other issues that do relate to safety, but not to this study are:
- Driving/shunting in areas that are not controlled by centralised traffic control (since this is another process, there will be different accident mechanisms, but a similar approach to the one developed in this thesis should be applicable);
- There may also be safety issues in processes that are not directly related to train path provision, for example passengers falling in trains, getting caught in train doors. These are not taken into consideration here.

1.5.6 Thesis outline

As can be seen from Figure 1.3, this thesis has the following structure:

Part I: Basic concepts and domain knowledge of railways, their organisation, and their technology will be explained in Chapter 2. Chapter 3 gives an overview of existing models in safety science in order to understand accidents.

Part II: A new safety model, the 'safe envelope of operations' will be developed in Chapter 4. It makes clear within which boundaries the operational process must remain. Chapter 5 then serves to introduce knowledge from control theory into the safety model. Thus, a cocktail of methods and models will serve to analyse the railways from several coherent perspectives. Chapter 6 will summarise these. Also, Chapter 6 will provide information on data collection for the next part.

Part III: In Chapters 7 to 9, the railways will be analysed. Starting point for the analysis is the train control process, which is performed by the driver. Then, the traffic control process will be considered, and finally the analysis will consider how design and management choices influence the abilities of drivers and traffic controllers to perform their tasks.

Part IV: Given the analysis from the previous part, suggestions for improvements will be formulated in Chapter 10 and conclusions will be drawn in Chapter 11.

[3] An exception would be if a train path at relatively high speed is effectuated alongside a busy platform. In this case, the train traffic may impose a direct risk to passengers, e.g. when they fall off the platform or when air turbulences are high.

Depending on the focus of your interest, there are several ways to read this thesis. In principle, every chapter can be read on its own; also, the thesis should be comprehensible if only the introductions or conclusions of every chapter are read.

2 Basics of railway organisation and operation

2.1 Introduction

This chapter explains some basic concepts of Dutch railway operation and control: it is written from a functional instead of the technical perspective which is common in a rail environment. This chapter is thus focussing on questions such as 'which tasks must be performed? What are their functions within the bigger system they are part of?' and not so much on the question how such a function is being performed in detail or on the technical working of a machine.

Although the focus of this dissertation is not on the organisational level but on that of the operational processes within the organisation, the organisational level must be briefly explained in order to understand the operational processes. It will first be explained which organisations are involved in the Dutch rail sector. This sets the scene in which the individuals within those organisations have to work. The focus will then move from the organisational level to that of the key players in railway operation. Their tasks will be explained, together with the most important tools that they use in their work: the traffic management system and the interlocking for the traffic controllers, the signalling system and train protection system for the driver, and the telecommunication system for both of them.

The information provided in this chapter will be necessary in Chapters 4 and 5 to understand examples that illustrate the theoretical framework developed in this dissertation. The information is also necessary to understand the information provided in Chapters 7 – 9, which will describe how the processes that are introduced in this chapter function in practice.

2.2 Players and markets in the railway sector

2.2.1 Railways as a service supply chain

A common framework to describe transport systems is the TRAIL Layer Model which is depicted in Figure 2.1 (Schoemaker et al. 1999). The basics of the TRAIL Layer Model have been further extended by Schaafsma (2001), for example to also include management of the means of production. The basic version consists of three layers: economic activities, transport

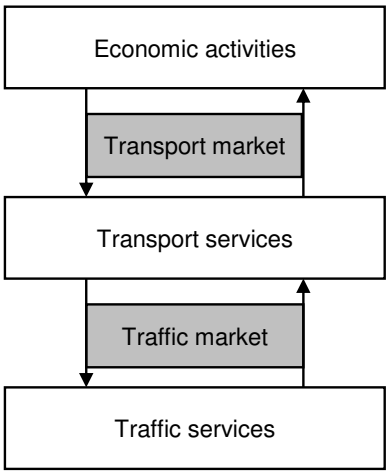


Figure 2.1: The TRAIL Layer Model (simplified).

services and traffic services. Since people perform activities in different locations, they need to make trips. This is a demand for transport services, which has a certain pattern over time and space. The transport services layer provides a supply pattern (e.g. moving empty seats). On the transport market, the supply and demand of transport services are balanced. If you buy a train ticket, you are active in the transport market. Similarly, suppliers of transport services need space on an infrastructure network for their vehicles to run on. This results in a demand for infrastructure which has also a pattern in time and space. Examples of this from outside railways are airlines that need slots on airports, but also cars that use a toll road. Also railway companies have to pay a fee for using the infrastructure.

EU-directive 91/440/EEC indicates that '*the future development and efficient operation of the railway system may be made easier if a distinction is made between the provision of transport services and the operation of infrastructure*'. Thus, the distinction between the two markets from the TRAIL Layer Model is compulsory in railways. A fee for infrastructure use '*shall be calculated in such a way as to avoid any discrimination between railway undertakings'*.

Although it is allowed that one organisation is active in both markets, financial accounts must then be separated. The railway industry can therefore be described as a service supply chain, similar to most other transport modes. Infrastructure providers and train operating companies both provide a service in time/space, comparable to a hotel allowing its guests to use a room for a certain while.

The TRAIL Layer Model is useful but it also has some drawbacks:
- It is not clear which actors provide the services.
- The layers are actually the same as the services that are exchanged on the respective markets. But it is not the transport service itself that offers a supply to the transport market, but its respective transport company. It is suggested to change the layers so that not the services, but the actors asking and providing these services are shown.
- Since transport is a service on an individual basis, it is possible to obtain 'transport' on the transport market. However, it is not possible to buy 'traffic' on the traffic market. It would be advantageous to call this the 'path market' as is done for example by Achilles (2008). A path is a trajectory in time and space in which space on the infrastructure is available for a certain user, and can therefore be obtained by this user.
- One could extend the model with one more level, since the infrastructure provider must also have its infrastructure provided by contractors.

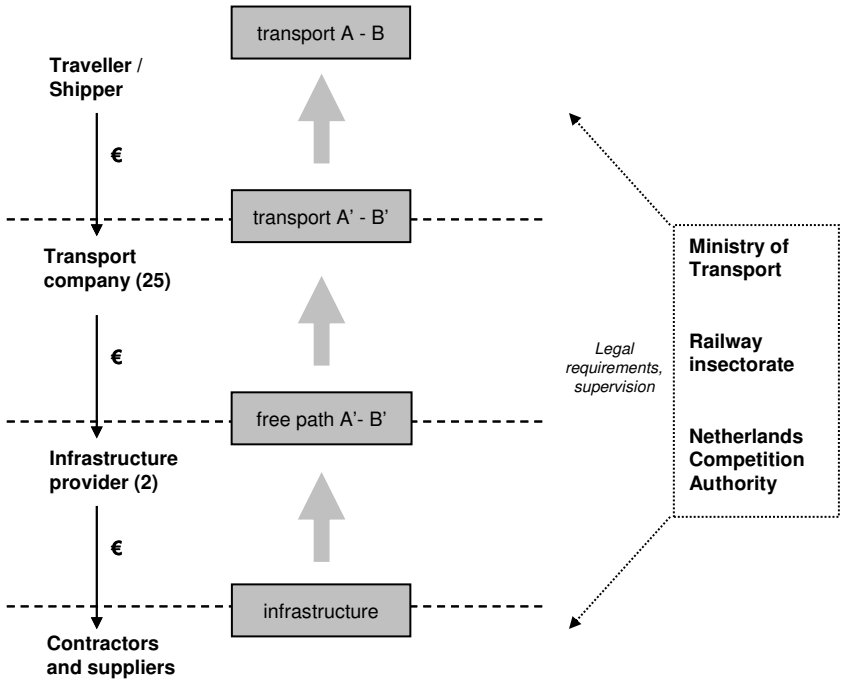


Figure 2.2: Service supply chain and actors in the railway system.

The actors have to perform a certain production process in order to provide the services, so that it is useful to swap the boxes and the arrows of the TRAIL Layer Model. The result is the service supply chain as depicted in Figure 2.2. Please note that several services from infrastructure providers and transport companies may be required in order for the traveller to 'assemble' a journey from his origin to his destination, for example when the traveller uses a bus service and subsequently two train services. For the convenience of the traveller, such combinations of transport services should be easy to connect, not only in time and space, but for example also in their fare structure. In other words, a transport service is a half-product, and its quality can only be assessed in relation to the overall goal of the traveller and not only on the basis of properties of the service itself. Since hardly anyone lives and works at a public transport stop, the services can only be provided to a point close to the origin and destination as desired by the customer. This is denoted by A' and B' in Figure 2.2.

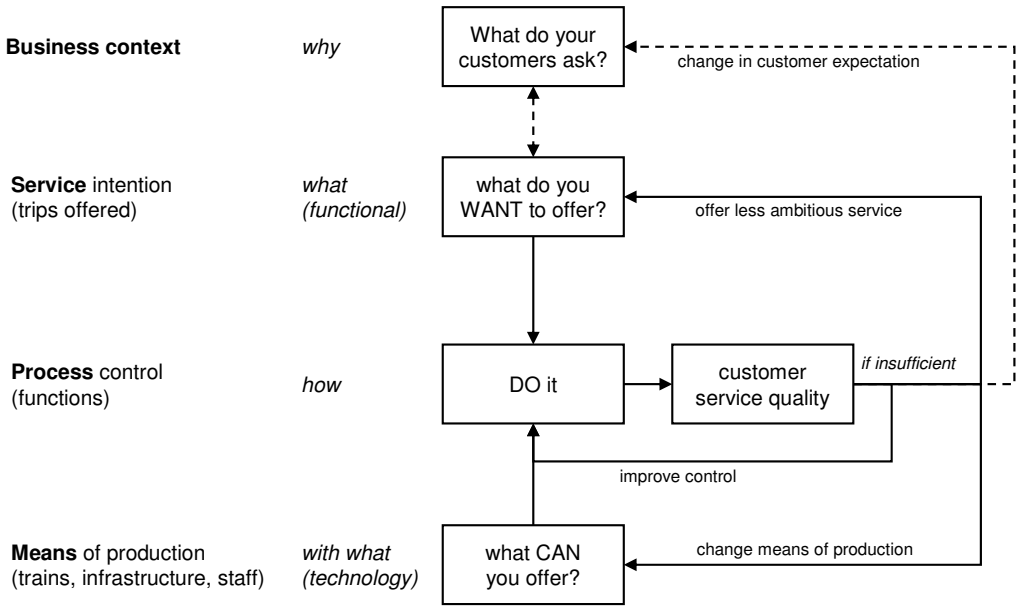


Figure 2.3: Services, processes and means in a business context.

The service supply chain may also contain other services. These include for instance: provision of rolling stock, provision of personnel or infrastructure maintenance. These are all activities that can be performed by the company that needs them. An example is a transport company that owns its rolling stock and that employs its own personnel. These can also be provided by an external supplier, such as a rolling stock leasing company or an employment agency. These services are generally not modified during the production processes[4] of transport services and train paths and therefore fall outside the scope of this thesis.

Each of the production processes in Figure 2.2 can be further split into steps, which are depicted in Figure 2.3. A similar stepwise approach was identified by CapGemini in their Integrated Architecture Framework, from which also the characteristic questions 'Why', 'What', 'How' and 'With what' are taken (Mulholland and Macaulay 2006). If the output is not as desired, changes can be made to one, two or all of the service offer, the means of production, or the way in which the means of production are used to supply the service.

1. Think of the business(es) in which your company wants to be active, depending on what is being asked for and what the company's capabilities are. *Why* is the company on earth?
2. Describe a clear goal of *what* has to be produced and for whom (customer interface).
3. Think of *how* this will be produced at an operational level. Which problems will operators face and what do they need to solve these?
4. *With what* can the production process be supported by better designs of timetable, technology, or by better regulations?

Technology can be used to make certain things possible. It is a first requirement then to know what these things precisely are and which function they must fulfil.

2.2.2 Transport services

Anno 2009, there are 25 transport companies active in the Dutch railway sector, of which 10 provide public transport services and 15 are active in freight transport (ProRail 2008b). In order to be able to offer a high quality of service, it is important in the first place to know which criteria travellers use to choose their transport mode. For freight transportation, the main criteria for shippers to choose a transport mode are (de Wit and van Gent 1996):

- Prices
- Transit time
- Reliability
- Capacity
- Timeliness

Vaessens et al. (2008) studied which criteria are of importance for passenger transport and related these to Maslow's pyramid. They are in the following list, the most important criterion being on the bottom:

- Experience (emotions)
- Comfort (physical effort)
- Ease of use (mental effort)
- Short point-to-point travel times (time)
- Safety, security and reliability (trust)

For a more elaborate list, refer to EN13816 (2002) or Veeneman (2002). Safety, reliability and travel times have a clear link to the traffic process control and clearly fall within the scope

[4] Although in the case of large disruptions, a main bottleneck is to reschedule staff quickly enough.

of this thesis. The other criteria (ease of use, comfort and experience) will not be taken into consideration in this thesis: their importance decreases and it involves much more than operational process control to achieve these.

Safety, security and reliability are in the basis of the pyramid, which means that these are basic customer needs: if these are not met, the transport mode will not be chosen, no matter how well it performs on the other criteria. The next-most important criterion is travel time. The modal split between public transport and private transport is strongly dependent on the ratio between their travel times. Research showed that if public transport travel times are at least twice as long as the travel time by car, it has no added value compared to alternative modes and only public transport captives (e.g. people without driver's license) will use these transport services; this forms about 20% of the market share. If public transport is as quick as the private car, it will gain a market share of approximately 60%. This means that 40% of all passengers never choose for public transport (Bovy and Van Goeverden 1994). However, public transport is as fast as the car on only 1% of all trips. Since equal travel times are difficult to achieve, and travel times that are twice as long have no added value, a rule-of-thumb is that public transport travel times should be at maximum 50% longer than the travel times by car in order to be a serious option for a reasonable number of travellers (Vaessens et al. 2008).

Customers compare whether the delivered services meet their needs. They can do this twice in public transport. Before deciding to make a trip by public transport, they can compare the market offer as described in the timetable to their demand; then after their journey, they can compare whether the service was delivered as it was promised (Pessers 2007). This means that it is of most importance not to deviate too much from what the customer expects once he has been informed about an arrival time at his destination station. Furthermore, whether these criteria meet acceptable levels depends not only on the services themselves but also on their environment: the acceptable price level, acceptable level of reliability and acceptable travel times depend on the levels that are achieved by comparable services; this is a normal market mechanism. Similarly, the acceptable level of risk is defined both by the individual's subjective assessment and also by the society.

2.2.3 Train paths

In addition to the 25 transport companies, there are several maintenance contractors and historic railway organisations. Although these do not provide transport services, they also need to move trains from time to time. Every day, the Dutch rail network serves 1.2 million passengers (for which 5,400 train runs are operated) and 400 freight trains transport 100,000 tons of freight per day. These trains operate on the Dutch network, which has a length of nearly 3,000 kilometres and consists of 6,800 track kilometres (ProRail 2008b). Each of the trains needs a so-called *train path.* A path is a trajectory in time and space in which infrastructure is available for a certain user. This can be difficult to imagine, since it is not only intangible but also a half-product. It is the main task of an infrastructure provider to allocate such train paths to trains, similarly as a hotel which allows its guests to remain in a certain room for a certain while. In the Netherlands, the network-wide infrastructure provider is ProRail. An exception is the BetuweRoute, a freight-only line from the Rotterdam harbour to Germany, which is operated by KeyRail. All train-operating companies are required to have a yearly contract with ProRail in which track access is regulated; they also pay a track access charge.

The term 'train path' is mostly used during timetable design. As soon as the infrastructure reservation is 'materialised' for real-time operations, the term 'route' is more common: e.g.

when switches are set into the right position for an approaching train, a route is being set. Strangely enough, no document is known to the author which describes the characteristics of a 'good train path' or a 'good route'. If that is so, quality of service will be difficult to measure. One could think of different quality aspects such as speed, desired arrival and departure times, priority service, rescheduling flexibility, type of safety systems provided and traction power supply. Similar to transport services, also train paths are a half-product. One of the most important quality indicators of a train path should therefore indicate whether it facilitates the transport service in a proper way.

2.2.4 Role of the government

The government is involved in this transportation system in several ways. By issuing legislation, it prevents the actors in the railway sector from behaving in a societally undesired way. Also, the government regulates the market via concessions. One concession regulates the conditions under which ProRail has to exploit the complete rail network, whilst another regulates how the largest train operating company, NS, should exploit transport services for passengers on the main parts of this network. Both concessions run until 2015. Besides NS, there are smaller train-operating companies which provide passenger transport for some regions; this is regulated through decentralised concessions. The market for freight train-operating companies has been liberalised, so no concessions are involved.

When the railways had just been split up, the governmental concessions were mostly used in a prescriptive manner to guide the behaviour of ProRail and NS. Since 1 January 2008 both companies are merely assessed on their performance indicators (see below). Since 1 January 2008, the Ministry of Transport uses an 'output oriented policy' to assess both ProRail and NS Reizigers. In contrast to what might be expected, this does not mean that both companies are from now on only assessed on their output, being the transport services and the train paths as shown in Figure 2.1. Instead, Camiel Eurlings, the Dutch Minister of Transport, explains that '*the output oriented policy means a new step in the relationship between NS and ProRail on one side and the Ministry of Transport on the other side. Via the transport plan and infrastructure management plan (both required to obtain the respective governmental concessions; added by author) I will agree with NS and ProRail on the performance to be delivered. Point of departure of the output oriented policy is that NS and ProRail, as professional organisations, are capable to estimate which performance will be achievable and to state this in their yearly plans. Once these agreements are made, NS and ProRail have to keep to their promises; if they fail to do so, I will enforce it'* (Eurlings 6 april 2009). The output-oriented control policy thus does not make sure that the railway sector is assessed on its output; even more importantly, it leaves the responsibility for what the output will be at the level of the sector instead of at the Ministry. Thus, it is still not made sure that the governmental policy of 'better utilisation' will be picked up.

2.2.5 Key performance indicators

The government uses key performance indicators (KPIs) to monitor the performance of ProRail and NS Reizigers. This section provides an overview of those KPIs that are relevant for this research, which means that they should have a relation with real time traffic control. For a full overview, refer to the 'Beheerplan' (Infrastructure Management Plan) and 'Vervoerplan' (Transport Plan) which are provided yearly by ProRail and NS.

NS Reizigers

For NS, the relevant KPIs are:

- Train arrival punctuality in 35 major nodes. Trains with anything up to 2:59 minutes delay are considered as punctual; NS CEO Hans Huisinga was forced to resign in 2002, because

the train punctuality in 2001 was only 79.9%, where 80.0% was agreed upon. Punctuality is the KPI that receives most attention. As of 2010, the threshold of lateness is to be set to 4:59 minutes, formally to harmonise the measurements with other European companies – the higher percentage trains that will be defined as being on time will be a welcome side-effect (Maartens 3 July 2009).
- The customer satisfaction with punctuality is measured separately.
- Number of realised connections between train services.
- Percentages of cancelled services.

These KPIs did not change during the research, but the values they should achieve did, so that continuous improvement was strived for.

NS Reizigers' main service is to transport passengers from A' to B' conform the goals that have been mentioned in section 2.2.2. However, no performance indicators have been set on short travel and waiting times or on point-to-point travel time reliability, even though these are crucial factors in determining the attractiveness of the offered services and are also identified as such in the White paper 'Mobility' (Ministerie van Verkeer en Waterstaat 2004a). Train punctuality is a misleading indicator as it does not minimise train delays (Al-Ibrahim 2008). Instead it will lead to concentrating delays as much as possible on one train that has already passed the 3-minute-threshold. Furthermore, it focuses on the punctuality of the trains rather than the punctuality of the passengers using them – which is not necessarily the same if passengers have to connect to other transport services. Although a separate KPI exists for realised connections, only a few connections (basically only between hourly operating intercity services on main stations) are measured by this indicator. As a result, most connections that are considered connections from a traveller's viewpoint (because they must change trains) can be broken and this will not be reflected in a reduced value of this KPI. Since it can lead to a higher punctuality, breaking such connections even appears to be beneficial when only the KPI values are regarded. Also, if the number of, or waiting times for scheduled connections get worse, such as in the 2007 timetable, no single KPI reflects this.

Current KPIs provide information about the production process that leads to transport services. As such, these can be useful indicators within NS Reizigers, but they do not measure the quality of the services provided. Also NS acknowledges that its current indicators focus more on logistics than on customer value. The company will therefore combine the KPIs arrival punctuality, realised connections and cancelled trains and weigh them according to the number of passengers affected. This will be a new indicator 'passenger punctuality' (Nederlandse Spoorwegen 2009b). Although this is a step in the right direction, it does not take away the criticisms mentioned before, because the new indicator is based on the same indicators that are currently in use; instead of being derived from what the customer needs during his journey.

ProRail

The situation for ProRail is similar to that for NS Reizigers, since, when this research started, no performance indicator was set on the company's output: the number and the quality of the train paths delivered to train operating companies (ProRail 2005b). However, ProRail's KPIs have been changing and are still in development; they are set out as far as possible in Table 2.1. Broadly speaking, ProRail's focus has shifted from an Infrastructure Manager to an Infrastructure Provider, so that also the focus changed from having a piccobello infrastructure to providing excellent train paths. This can best be clarified by the following example: the availability of the rail network was measured by multiplying the number of 'train service affecting irregularities' (TAOs) by the average repair time (FHT) (ProRail 2006). The

Table 2.1: Relevant KPIs for ProRail (ProRail 2005b; 2009a).

Desired result	KPI 2005	KPI 2009	description
Customer appreciation		Customer satisfaction*	Customer satisfaction about ProRail's services
Rescheduling quality (Traffic Control department)	Route setting irregularities*	Route setting irregularities*	number of times that a route was not set when it should have been according to the (updated) plan
	Rescheduling according to rules	Rescheduling according to rules	Percentage of schedule changes that was performed as agreed upon with train operating companies before operations started
		Train path	Indicates how well ProRail can deliver train paths
		Resilience: alternatively provided train path	Indicates how well ProRail can continue to deliver train paths even when the original schedule is disturbed
Non-discriminatory capacity allocation (Capacity Management department)	Successful appeals at the NMa (Netherlands Competition Authority)		Indicates how often ProRail does not distribute capacity transparently between train operating companies
		Capacity allocation train path	Indicates which percentages of requested train path could be incorporated in the timetable
Infrastructure availability and relaibility (Infrastructure Management department)	TAO		measures the number of times that infrastructure is unavailable due to a irregularity
	FHT		time between detection of a TAO and the moment that the infrastructure is available again
	TVP		time window in which planned infrastructure maintenance is performed; TVPs are differentiated in the level of traffic disruption they cause: none; some; much; extremely much
		Punctuality	Makes ProRail also responsible for the punctuality on the Dutch rail network
		Non-availability (maintenance)	Percentage of time that infrastructure is not available due to scheduled maintenance
		Non-availability (failures)	Percentage of time that infrastructure is not available due to unscheduled maintenance
Safety (safety department)	Safety index*		Index number indicating whether the safety level went up or down
		Collisions	Number of collisions
		Derailments	Number of derailments

* indicators indicated with an asterisk were shown for information only and were not assessed formally by the Ministry of Transport.

problem with the number of TAOs is that these do not take the impact (duration and affected area) of an irregularity into account. Although the average repair time FHT compensates more or less for the duration, it still does not indicate how many train paths have been affected during that period. Therefore, another multiplication factor was introduced that represented whether an 'important' or 'unimportant' line section is out of service. This was quite a difficult way to estimate how many train paths could not be delivered. These KPIs have since been abandoned in favour of train path related indicators.

Table 2.1 provides a broad overview of relevant KPIs during the time period of this thesis, reflecting the situation in 2005 and in 2009 (ProRail 2005b; 2009a). Again, only KPIs relevant for this research are shown. The 2005 KPIs are related to the single departments of which ProRail consists, which was clearly expressed by a staff member of a train operating company: *'ProRail is a three-headed entity with three objectives. They find operating trains troublesome, since it leads to track wear and the trains are also in the way when the tracks must then be maintained'*. The 2009 KPIs reflect ProRail more as a whole and focus more on train paths and customer relations.

These KPIs are still changing:
- It is planned to have abandoned the KPIs 'route setting irregularities' and 'rescheduling according to rules' before the start of 2010. They will be replaced by the new KPIs 'train path' and 'alternatively provided train path'.
- KPI Punctuality will be replaced by a KPI train delay minutes, as is usual in other countries. Train delay minutes can then be attributed to those organisations that caused them.
- A KPI 'Modal Split' will be developed together with Rijkswaterstaat, the Dutch highway authority. It is meant to measure whether rail transport is an attractive alternative to the private car.

With its new performance indicator 'Train Path', ProRail appears to be on a track where the company's fundamental *raison d'etrê* is being translated into a performance indicator. ProRail plans to use this KPI both on a management level (to assess the company's performance as a whole) and on operational level (to assess the effectiveness of its production processes).

In conclusion, several KPIs are now defined on a one-company-level. However, as stated in sections 2.2.1 and 2.2.3, it is also necessary to assess train paths and transport services on their ability to fit into 'the greater good' of the customer's journey, similarly as one can only assess whether grain is of sufficient quality once it is known whether it will be used for forage or for a pie. No such KPIs on the system level have been identified, which means that even a total fulfilment of all KPIs does not necessarily lead to attractive public transport services. A fundamental dilemma here is the question whether it is acceptable to assess a company on an indicator which it cannot be fully responsible for, because it can influence its value only partially (*Vervoerconcessie hoofdrailnet* 2005). However, not making the companies responsible for the overall outcome risks that misfits between responsibilities of single companies lead to problems at the (customer's) top level. Making them jointly responsible would also encourage collaboration between all those organisations which have some influence.

In a discussion with one of ProRail's responsible employees it became clear that it was regarded as a drawback that customer satisfaction is a subjective measure. However, quality is defined as the degree to which a set of inherent characteristics fulfils requirements (ISO9000 2000) and these requirements are always subjective. It is argued by Fay (1999) that high quality products and high quality services can *only* be judged as such when they do fulfil

customers' (subjective) needs, whilst Rietveld (2005) showed that supply-oriented quality indicators are systematically biased and overestimate performance quality. Also NS indicates that: '*for NS, customer satisfaction is the most important indicator for our quality of service*' (Nederlandse Spoorwegen 2006). Subjectivity apparently is inherently interwoven with customers' service perception, so that it is not at all bad to include subjective measures in KPIs.

The remainder of this thesis will focus on the quality and safety of the train paths that are delivered, which overlaps with the KPIs for ProRail Traffic Control. Although it is outside the scope of this thesis, a final note is that all safety indicators are related to train-related accidents (derailments and collisions) or maintenance-related accidents. Since ProRail has also a task to provide transfer space (e.g. on platforms) to passengers, a safety indicator should also be developed for accidents that may occur to passengers when they change trains.

2.2.6 Safety requirements

Safety targets are set in a policy document (Ministerie van Verkeer en Waterstaat 2004b), of which the most important notions are that:

- the safety level should remain at least on the current level: a maximum of 0.15 fatalities per 10^9 passenger km; in practice, more attention is given to the more concrete targets of the SPAD workgroup (see section 2.2.7);
- the whole network should be equipped with some form of automatic train protection;
- a registration and monitoring system should be in place in order to provide an overview of the safety situation.

It can easily be seen that these are sector-wide targets. In the European research project SAMRAIL it was indicated that it is necessary to measure the overall level of safety of the railway within an European member state first, and then see how the performance of individual companies relates to that (El Koursi et al. 2006). This gives opportunities to debate how such nation wide targets should be distributed over the different independent companies. However, too much splitting up of risk targets per company will make it likely that companies will not cooperate, since they are all responsible for their 'own' partial KPIs. However, co-operation is a necessity, not an option, and needs to be represented by some system-level KPI.

Safety requirements are defined in the railway law and its underlying regulations. The Dutch railway law was recently rewritten and came into service in 2005. It replaced the old railway law which stemmed from 1875. The law contains several underlying regulations, for instance with regard to the formal meanings of signal aspects, prescribed maximum braking distances, required on-board equipment and staff education. Many of these requirements are set on a relatively coarse level and must be further detailed in company-internal rules, such as infrastructure design rules or company-internal education. Also, it delegates several responsibilities to ProRail or other parties. The signalling norms are an example of the latter. These can be supposed to function as a 'manual' for infrastructure design, but are also used as a check for that design. This is especially true now that external engineering firms do much of the design work for ProRail. When the design fulfils the norms and standards, it is generally considered as a sufficiently good design.

2.2.7 SPAD workgroup

Following an increasing number of train collisions, the Ministry of Transport formed a sector wide SPAD-workgroup, so that the problem of trains running through red signals could be solved in a coordinated way among all relevant actors. This workgroup was formed in March 2004. On 21 May 2004, there was a relatively severe collision between two trains in Amsterdam. Although there were no deaths, 18 passengers were injured and rail traffic was

blocked for three days. The Dutch parliament asked for measures to prevent trains from running through signals at danger and the sector-wide SPAD-workgroup was granted €40 million (later increased to €55 million) by the Ministry of Transport (Ministerie van Verkeer en Waterstaat 29 november 2004). The SPAD workgroup aims to:

1. Reduce the risk associated with SPAD by 75% in 2009 relative to the 2003 level. The risk is calculated by assigning points to every signal on several criteria (Wright and Vorderegger 2005).
2. Reduce the number of SPADs by 50% in 2009 relative to the 2003 level.
3. Increase safety on the 100+ locations before 2007, prioritised on probability, consequences and risk (100+ locations are situations in the network where a train that runs through a signal at danger can collide with a train running faster than 100 km/h). This was mainly a politically driven desire.

The workgroup started with four projects (Schmeink and Beuk 2006):

- Analysis of station layouts (signal visibility and use of routes);
- Drivers program, consisting of a vigilance test, instructions to improve self-awareness, and instructions for use of communication means in the cab;
- Route setting instructions to prevent the automatic route setting system from automatically setting routes that have frequently been SPADed;
- Development of a patch to the ATB-system so that it would also stop trains when they were about to run through a red signal (ATB-VV, a Dutch acronym for ATB-improved version), which was installed at approximately 1,150 of 9,800 signals and came into operation in the first half of 2009.

In 2007, the following projects were added to the program because the expected reduction in SPAD numbers was not yet visible (Verheij and van Dijk 2008):

- Add snow and sunlight hoods to prevent signal lights from being obscured;
- Coordination in the driver's program, because every train operating company applied it differently;
- Research into effects of signalling failures which cause SPADs when a signal shows red without prior announcement;
- Collection of data after every SPAD incident;
- Analyse influences of different planning procedures;
- Research into SPAD incidents by freight trains, as these were relatively frequent;
- Research into SPAD incidents by shunting trains, as these were relatively frequent;
- Research into SPAD incidents caused by rolling stock that was parked incorrectly and rolled through red signals;
- Extensive research into 10 multi-SPAD signals to find underlying causes.

In relation to the cascade model, which will be explained in section 2.3.1, the scope of the SPAD workgroup was first mostly limited to interaction between drivers and signals (signal visibility, driver vigilance) and automatic train protection. In a later stadium, also more research projects were added. These also take other functions, such as timetabling and route setting into account and aim to provide more understanding of why SPADs occur.

2.2.8 Inspectorates

There are also two supervising bodies that may intervene in the market mechanism. The railway inspectorate focuses on safety and has the following tasks (Inspectie Verkeer en Waterstaat 2007g):

1. Supervise that railway companies comply with rules and legislation. Such inspections each consist of three main activities: retrieve information, interpret this information, and intervene if necessary.
2. Grant permissions for those activities that require one;
3. Since inspections result in knowledge, a third task is called 'knowledge, advise, reporting'. This helps to improve inspections, find trends and developments and to form new policies if necessary.

The Netherlands Competition Authority (Nederlandse Mededingingsautoriteit, NMa) judges whether the competition rules between different players in the market are violated. A third organisation is an independent accident investigation board, which can analyse severe accidents, or patterns of lesser incidents, and thus oversees all other actors, including the Ministry and the inspectorates.

Now that the organisations that make up the Dutch railway system have been described, the next section will focus on several functions and actors that are active within these organisations to produce the services as required.

2.3 Control processes in railway operation

2.3.1 The cascade model

In order to provide transport services, the transport companies and the infrastructure providers must co-operate: the train path and the transport service have to be produced simultaneously. Traditionally, the rail sector has been hierarchically organised to ensure this co-operation. This process can be depicted in the following model (see Figure 2.4), which we have christened the 'cascade model'. In it every actor is dependent on the decisions that have been taken before: the degrees of freedom decrease as the moment of operation comes closer. A new or modified infrastructure has to be planned years in advance. Also timetable planners meet years in advance to decide what operators should do in real time. Operators are thus restricted in their possible decisions. They cannot use tracks that have not been built and they cannot deviate much from the timetable once it has been published. Finally, at the bottom of the cascade, is the train driver who has hardly any other degree of freedom apart from controlling the train's speed. This working method has existed almost as long as the railways themselves, as is shown in many examples from the literature (Rolt 1966; Jongerius 1993). In the past, the cascade model was necessary as real-time communication was not possible in the early days of the railways. The cascade was meant to ensure that all workers – whether on one of the trains or at one of the stations – were working by the same reference and the same rules, so that the operational process could run smoothly. In order for this to work, the organisational culture became militaristic. Orders were orders and a precise execution of the plan had to be achieved (Rüter 1960). This leads to the question whether this is still the best way to organise operations: given today's possibilities for real-time mobile communication, there appear to be other, maybe better ways of ensuring that all actors work on the same goal.

In the cascade model, the level of abstraction is high on the top-left side and low on the bottom-right side: train drivers talk about a specific signal and a specific train movement, where system architects talk about signals in general and timetable planners often consider a whole series of trains (e.g. the same pattern every hour, all year long). Following the argumentation in section 2.2.2, it is also possible to include quality of service in the cascade model, since it includes the customer needs (coming from the top as an external goal or boundary condition), the service offer as promised and the actually realised service.

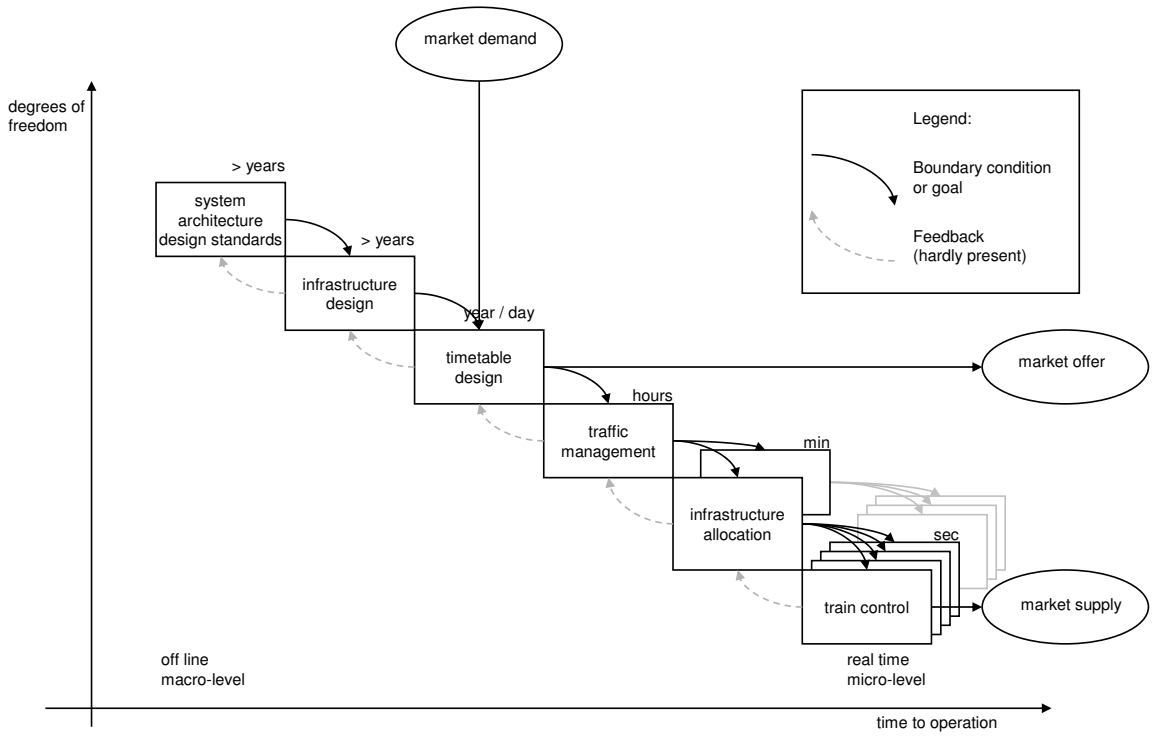


Figure 2.4: The cascade model.

2.3.2 Railway operation as control loops

Recent insights describe railway operation as nested control loops (Fay 1999; Lüthi et al. 2006; Van den Top 2006; Oetting 2008). These ideas are applicable to the cascade diagram as well. In general, one could say that each actor should perform a Deming circle (plan – do – check – adjust), as is shown in Figure 2.5. Although in reality this is not always the case, this still can be a useful model as it will then show that certain functionalities might be missing. Because there is a hierarchy in decision-making, the action of a first actor sets a goal or a condition for the next actor, who performs another Deming circle. Both actors can adjust their own processes (internal Deming circle). A lower-level actor can also give feedback to a higher-level actor if conditions are too strict or too lax. For the higher-level actor this can be seen as an external check of his activity. The next sections will describe the different functions in the cascade model.

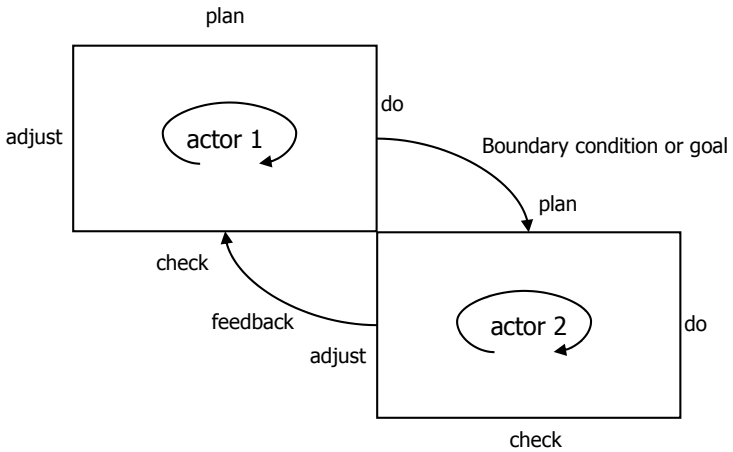


Figure 2.5: Two control loops.

2.3.3 System architecture and design standards

The function of system architecture and development of design standards has two objectives:

1. Translate requirements to the system (e.g. legislation, customer demand) into more precise requirements for subprocesses and subsystems.
2. Make sure that these subprocesses and subsystems work smoothly together such that the overall outcome is as required.

Thus, the task of system architecture is to arrange the systems elements in such a way that the desired functionality will be achieved. One of the means to do this is by posing design norms. Such norms can be developed once, so that not every sub process needs to re-invent the wheel.

Under the function of system architecture, we also include the managers and designers who decide on the realisation and design of tools and procedures that assist drivers and traffic controllers in their daily work, since these also contribute to smooth cooperation in order to provide the desired result. Examples are traffic management systems and driver information systems, but also operating procedures, such as the handbooks for drivers and traffic controllers (ProRail 2003).

Safety norms

As far as safety is concerned, the relevant rules for this research are dominated by the signalling system. One of the most important rules follows from the long braking distances that are inherent to rail traffic. Since braking distances are so long that a driver cannot reliably see and stop before obstacles ahead, he must always be warned explicitly at least at braking distance before any speed reduction should be achieved. It is the function of the signalling system to show (mostly safety-relevant) information to the driver. A signalling system is not so much a collection of physical objects (e.g signals and traffic signs) but rather a coherent and complete set of messages that can be given to the driver (e.g. also lineside signs, in-cab signals, gestures, or auditive warnings). In the Netherlands, the meanings of signals are defined by the Ministry of Transport and issued in an appendix to the regulations that are part of the Dutch railway law (*Regeling Spoorverkeer* 2004). ProRail's design standards for infrastructure and signalling provide a further level of detail and state, for instance, where signals can be placed (Algemene Voorschriften / Ontwerpvoorschriften, abbreviated AV and OV) (ProRail 2005a).

Service architecture

Also basic decisions about the services that will be offered can be considered as system architecture decisions. Examples are the types of train services that the company wants to offer, e.g. intercity services and/or local services. For freight transport, often a distinction is made between transport of containers, chemical substances, dry bulk or transport of wagonloads. Another important decision is the type of timetable that will be operated; refer to section 2.3.5 for more information about this topic. Since signal distances define the minimum headway between two trains and thus the minimum time interval between two successive train movements, these form a restriction for timetable design. In order to make sure that a timetable design is technically feasible, timetable design norms are developed based on minimum headways; these norms also contain some margin in order to keep the timetable robust against small disruptions. The timetable design norms are published in the Network Statement (ProRail 2009b).

Because the design norms are based on the capacity of the signalling system, fundamental changes to signalling can have impacts on all subsequent actors in the cascade diagram. Cab signalling does not only affect the driver's task, but it can and will also have influences on

timetable design, infrastructure design and on traffic control, since it provides much more functionality than current signalling systems. An example is the introduction of signalling in the driver's cab by means of the European Train Control System (ETCS Level 1, 2 or 3) (Eichenberger 2007; Winter 2009), which is obligatory on new railway lines by EU-directive 96/48/EC and is meant to gradually replace national signalling systems on existing lines by directive 2001/16/EC. A short description of this system can be found in Appendix A6.

Paradigm change on architecture level
The paradigm in the 1990s was to extend many track sections to four tracks and to build level-crossing free junctions, so that different train movements would not interfere. Both types of measures were main parts of the Dutch future plan Rail 21(Nederlandse Spoorwegen 1988). Nowadays this approach is normally considered to be too expensive and it is primarily aimed to gain some capacity by relatively small modifications, such as longer platforms, or signal relocation schemes, as was proposed in the subsequent future plan Benutten en Bouwen (Project Benutten & Bouwen 2003a).

2.3.4 Infrastructure design

The function of infrastructure design is to provide adequate 'hardware' for the train paths that are desired by train operating companies. When it is felt that the current infrastructure is not capable of providing these, infrastructure modifications may be appropriate. One should not only think of additional tracks, since also other measures are often used, such as increasing maximum speeds, changing the lay-out of a switching area so that more train movements can be handled in parallel, or signal relocation schemes to enable shorter headways.

Since large scale infrastructure modifications are expensive and have environmental impacts, such expansions are always subject to many analyses and debates, so that it takes several years between the first idea of a modification and its realisation. Often, infrastructure is built for a lifetime of up to 100 years: most railway lines are still in the same place where they were built in the 19th century. These long time spans make it difficult to make a precise prediction of the intended infrastructure use.

2.3.5 Timetable design

Rail traffic is precisely planned in a timetable. The timetable serves two functions (Weits 1997; Schaafsma 2001):
1. Public function: a catalogue of services offered to customers;
2. Working function: a planning of which means of production (infrastructure, rolling stock, staff) is needed at which moment.

In some countries these two functions are clearly separated, e.g. in Great Britain where a distinction is made between the public and the working timetable. The public timetable generally indicates departure and arrival times of train services in minutes, where the working timetable is often more detailed, for instance by giving also passing times at junctions or being precise in half or quarter minutes. In some cases public arrival times are later than the working ones. This gives the train operating companies some freedom to perform last-minute optimisations without affecting the promises to the customers. In the Netherlands, these two functions have not (yet) been made independent.

Public timetable
Different types of timetable structures can be identified (Maarschalkerweerd 1997; Hansen and Pachl 2008):
- Regular timetable, indicating fixed intervals between successive train services;

- Clockface timetable, indicating that the same pattern of services is repeated. The cycle time often is one hour.
- Symmetrical timetable, indicating that travel times in both directions are identical;
- Synchronised timetable, indicating that trains from all directions meet in major stations to offer connections to most or all other trains.

In 1938, the Dutch were the first to come up with a regular, symmetric and clockface timetable. This type of timetable is easy to remember for travellers and the original design evolved more or less accidentally to a synchronised timetable. This timetable was further established in the 1970s in the future plan Spoorslag '70 (Nederlandse Spoorwegen 1969). This timetable served as the basis for all timetables in service until December 2006, when a new structure came into service (Avelino et al. 2006). Depending on the type of timetable that will be operated, different requirements are posed on the physical railway network. For instance, to operate a clockface symmetrical timetable, the infrastructure must provide meeting points for trains in opposite directions at time intervals of half the train frequency. If a synchronised timetable is to be operated, these meeting points must be in hub stations, which then also poses requirements on the running times between such stations.

Working timetable

In general, the timetable is conflict-free, which means that the intended train paths do not need the same track section at the same moment. In order to make sure that no track conflicts will occur, the blocking time theory is generally used, especially in German speaking countries. This theory was developed by Happel (1959) and is now also recommended to calculate capacity use by the International Railway Union (Union Internationale des Chemins de fer, UIC 2004). Hansen and Pachl (2008) provide a state-of the-art standard work on timetable design.

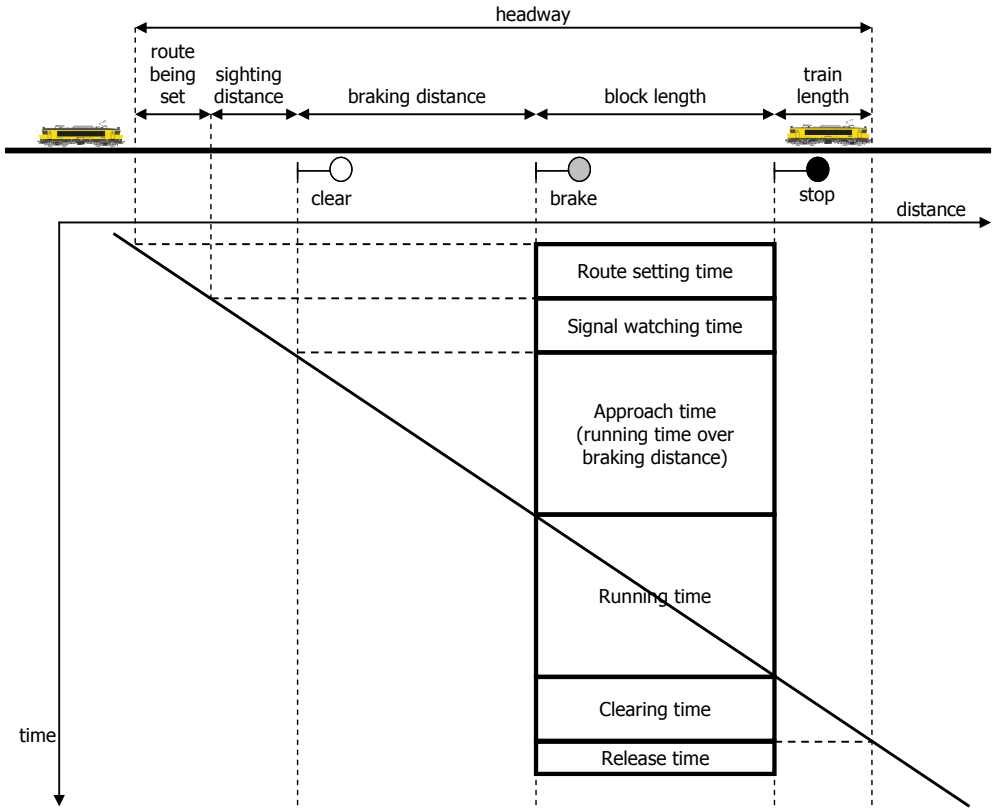


Figure 2.6: Blocking time model.

A blocking time diagram (Figure 2.6) in fact shows the intended or realised track reservations for every train; the figure shows the minimum headway for unhindered movement of two successive trains. Since a piece of track can only be assigned to one train at a time, the shortest possible (unhindered) headway will result when the second train is at sighting distance of the distant signal precisely at the moment that this has cleared, because the previous train has left the block for which this signal is a distant signal[5].

If blocking times of two trains overlap, there is a conflict in the plan. If this plan was to be executed, one of the two trains would meet a restrictive signal aspect because its route cannot be given. Timetable design is generally done in such a way that the timetable is conflict-free, so that no train will meet restrictive signal aspects if all trains are running on time. This means that a route should be cleared ultimately approximately 9 seconds before the driver passes its corresponding distant signal. An example of a blocking time staircase (subsequent reservations for a train movement) is displayed in Figure 2.7.

In the Netherlands, it is not common to make blocking time calculations during timetable design. Especially on tracks between stations, block lengths are very regular and almost never normative for the timetable design. Standard planning norms are used to determine feasible train headways. These are meant to check that a timetable is feasible and to ensure the design of a timetable that is robust during normal operations (Schaafsma and Weits, 1996). The timetable design programme DONS (Designer of Network Schedules) is used to find a timetable for a given number of services and connections between these, which takes the required headways into account (Hooghiemstra et al. 1998). The headway norms are based on a blocking time calculation for a representative Dutch situation, and include some buffer time to avoid that a small train delay immediately affects the next train. Actual blocking time calculations are only made in exceptional situations, e.g. when it is desired to deviate from the planning norms. Typical values for the Dutch situation are shown in Table 2.2.

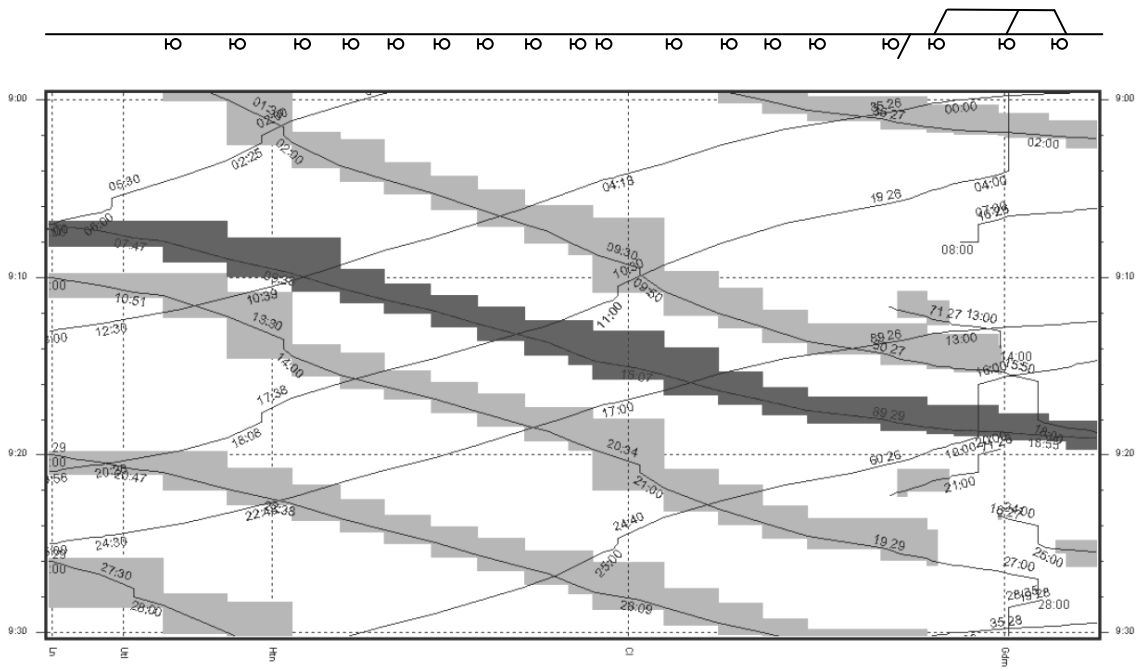


Figure 2.7: Blocking time diagram for one direction of a double track line (Weeda 2005; reproduced with permission).

[5] More detailed information about signalling can be found in Chapter 7.

Table 2.2: Typical blocking times for the Dutch network.

	Passenger	Freight
Block length	1500 m	1500 m
Braking distance (previous block)	1500 m	1500 m
Train speed	140 km/h = 38,9 m/s	100 km/h = 27,8 m/s
Train length	400 m	700 m
Route setting time (3 sec at block signals)	12 s	12 s
Signal watching time	9 s	9 s
Approach time	1500 m / 38,9 m/s = 39 s	1500 m / 27,8 m/s = 54 s
Running time	1500 m / 38,9 m/s = 39 s	1500 m / 27,8 m/s = 54 s
Clearing time	400 m / 38,9 m/s = 10 s	700 m / 27,8 m/s = 25 s
Release time	3 s	3 s
TOTAL Blocking time	112 s	157 s
Standard headway norm	180 s = 3 min	240 s = 4 min

A passenger train can be followed by another train after 112 seconds and a freight train can be followed after 157 seconds. With inclusion of buffer time these are rounded up to headways of 3 and 4 minutes respectively. Note that the timetable calculations are done on a level of decimal minutes, but these are rounded to whole minutes before the timetable goes into operation. Such buffer times allow for small delays, but when a train is delayed for some more minutes, it will miss its intended path and must be given a path at a later moment, which requires some spare capacity left. UIC Code 406 (UIC 2004) advises that no more than 60% to 85% of the available time should be allocated to train movements, depending on the type of line. Since the Dutch planning norms contain buffer time, the UIC advice is almost always met automatically. However, this is a very broad advise and it is generally advantageous to have a more precise allocation of spare capacity (Weeda and Wiggenraad 2005).

Especially in Germany many analytical and simulation models have been developed to estimate operations quality from a number of timetable characteristics. An overview is given by Hansen (2006), who also indicates that it would be worthwhile to use a stochastic approach to buffer time calculation between blocking times, in which measured train delay data are used to provide an accurate distribution of delays. This would allow to put the right amount of buffer time at the right places (Hansen 2001; 2004). However, in the Dutch situation, such detailed models are not used.

To get an impression of the dynamic performance of the timetable, a concept timetable can be simulated. In the Netherlands, this is done by means of SIMONE (Simulation Model for Networks) (Bouwman and Middelkoop 2001). In these simulations, the traffic is slightly disturbed and train punctuality is measured. Also several studies have been performed to find out how these buffer times (Goverde 2005) and running time margins (Vromans 2005) can be applied most effectively. This allows the making of statements about where critical parts in a timetable design are located and to give an oversight of how delays spread through the network, but these methods are not (yet) applied during timetable design.

The methods so far are all meant to assess timetable performance in advance. In 2006, a performance analysis office was established within ProRail, which analyses where conflicts during operations occur. This results in advice to tweak the timetable (Weeda and Hofstra 2008); several cases have also been found where strict application of planning rules leads to results opposed to what was meant to be achieved (Schaafsma and Weeda 2009). Up to this point, the timetable is still not more than a one-hourly pattern. When one of these concept timetables is selected for application, this pattern is fed from DONS into the VPT system (Vervoer per Trein), the planning and control software that is used for timetable design and for traffic management. The planning module of VPT allows planners to extend the pattern to

a real timetable, including trains that do not fit into the hourly pattern, first and last trains of the day etc. Once the timetable design results in a fixed design, it is given to all relevant actors, such as passengers (public timetable), traffic controllers, drivers, et cetera.

2.3.6 Traffic management

The rail system is highly dynamic. Disruptions are the rule rather than the exception. It is the task of ProRail's Traffic Control department to deal with those disruptions, which means a constant monitoring process and intervention if necessary. The principal task of rescheduling involves identifying and resolving conflicts which may arise during actual operations (Jacobs 2008).

Currently, the traffic control organisation is set up at two levels: regional and local traffic controllers. The tasks of regional traffic controllers are covered in this section; that of local traffic controllers in the next section. Their common objective is to decide which trains will make use of which parts of the rail infrastructure at which moment. If there are no disruptions and all trains are running according to their schedule, there is no need for traffic controllers to act. It is the task of four regional traffic controllers to make short-term changes to the timetable, using another module of the VPT system, called VKL (an abbreviation for Verkeersleiding). Examples are:

- find a path for a freight train that was ordered at short notice;
- change the order between trains in the case of delays;
- decide on back-up scenarios in the case of blocked tracks (degraded services).

In those cases in which a track or a line cannot be used, fall back plans ('versperrings-maatregelen') exist which prescribe which trains will continue to run and which will be cancelled. These documents are intended for use by operators to restore the timetable as soon as possible in the case of disruptions (ProRail, 2003). However, these plans do not try to optimise the usage of the line section with the temporarily reduced capacity, but they just make cuts in the original plan until no more conflicts exist. The performance thus is lower than it could otherwise be (Koopman 2007).

2.3.7 Infrastructure allocation

Route setting

Local traffic controllers determine which train will be given authorisation to use certain track sections at which moment. About 75 local traffic controllers (based in 13 local control centres) have the task to provide real-time infrastructure operation. The area under the control of one local traffic controller depends on the workload and consists in general of one line, one station or half of a big station. Local traffic controllers use yet another module of VPT, called PRL (Procesleiding, which means process control). Their workplace (Figure 2.8) consists of several screens showing the infrastructure layout. Train detection systems are used to detect which track sections are actually occupied by trains. Based on this information, the traffic controller and the automatic route setting system (ARI, automatische rijweginstelling) can decide which train needs a route next and assign it by clearing a signal. Track sections that are occupied by trains are indicated in yellow; routes that are set for trains are indicated in green.

Another screen shows a list of 'plan lines', each representing a route to be set for a certain train (this will be explained in more detail in section 8.2). Under normal circumstances, the automatic route setting system sets the switches and signals as defined by the 'plan lines' and the local traffic controller only needs to monitor the process. In the case of deviations from the timetable, the traffic controllers should adapt the plan that the automatic route setting



Figure 2.8: A local traffic controller's workplace.

system will execute. The timetable then serves as a reference for the actual situation and also as a prescription for the future course of events. As long as the trains are punctual and there are few deviations from the timetable, this approach works well. However, the timetable's prescriptive function gets less and less useful when the number of delayed trains increases, because the timetable is constructed on the assumption that all trains are running to time. The planned optimal order of trains is, in that case, based on outdated information (Van den Top 2005; Vorage 2006).

Similar to regional traffic controllers, also local traffic controllers have rules that tell in which way the planned order should be changed in the case of delays. These rules assume that only one train is delayed and the other trains in the network are running according to the schedule. Traffic controllers indicate in interviews held during this research that this is generally not the case. In undisrupted service all trains run to schedule and the rules are not necessary. In the case of disruptions, most trains have at least small delays and the rules do not work. The resolution of traffic conflicts by these rules becomes less effective just at the moment when it is most needed. If real-time communication between trains and traffic control centre about the train's actual position and speed and the expected moment of route availability were possible, the infrastructure usage could be used more efficiently (Hansen 2001).

Interlocking to prevent erroneous route setting

Despite the conflict-free timetable and the rules for disturbed operations, simultaneous claims on routes may arise. Of course, a route cannot be assigned to two trains at the same moment, as this would possibly result in a collision. Assigning infrastructure to trains is therefore a safety-critical task. All requests for routes, stemming from the traffic controller (or the automatic route setting system on behalf of him) are processed by an interlocking system. This is a vital safety system, which administers all units of the infrastructure, such as track sections and switches. If all programmed safety conditions are met, it sets the elements to the position necessary to form the route as requested. Only when all units in a route are in the correct state, will the interlocking clear the signal that gives access to the route, and so inform the driver about the ability to use the route that has been set. The route elements remain locked until the train has cleared them. The elements are then released and become available for the next route to be set. This is the same cycle as is represented by the blocking time model in section 2.3.5. Tracks between stations or junctions are often subdivided into so-

called block sections, such that a route for a next train can already be set and the next train can follow the first train before it has left the track by entering the next station. Block signals prevent collisions with the preceding train with a similar logic as an interlocking does in stations.

Communication with other controllers and drivers
When a traffic controller wants to change the planned order of two trains between two stations in a way different from the rules, he should contact the regional traffic controller to ask for permission. The local traffic controller is also the contact person for train drivers, for instance when they want to report track defects or need emergency services. For this communication, each workplace has a telephone set (at the right-hand side of Figure 2.8). There is a separate mobile network for communication with drivers, GSM-R (Global System for Mobile communications for Railways). This network is similar to commercial GSM networks, but it has some additional functionality. Examples are:
- Group calls, so that a traffic controller can reach all drivers in a certain region;
- Priority calls, so that a connection will always be made for an alarm message, and if necessary other connections will be broken;
- Functional telephone numbers, so that the traffic controller does not need to know all driver's phone numbers. He can reach a driver by selecting the relevant train number, which is also used for route setting and is therefore easily known.

2.3.8 Train control and automatic train protection

Controlling the train movement
Finally, the last function in the cascade diagram is to control the train, which is generally performed by the driver. It is his objective to bring the train timely and safely to its intermediate stops. By doing so, he finally executes the plan that has been already through the (virtual) hands of so many other actors. He has the plan in the form of a printout of the timetable for his train, or at some train operating companies it is available on a minicomputer. It indicates arrival, departure and through times at stations the train will pass.

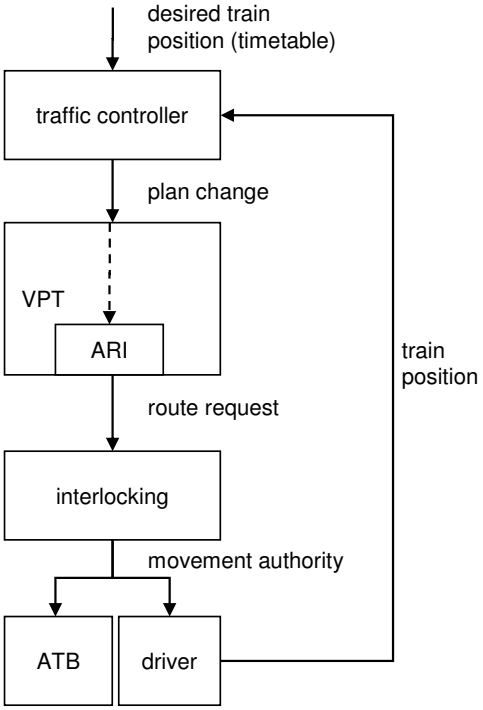


Figure 2.9: Interaction between traffic controller, control system, interlocking and drivers.

These processes of the traffic controller and the driver need communication as is shown in Figure 2.9. Under normal operations, the signalling system is used to inform the driver about permissions and orders – the movement authority – from the traffic controller. The signalling system is designed in such a way that signals are normally at red unless a route is set for a train. A driver can therefore only execute his task after the traffic controller has assigned a route to his train. The details of the signalling system are elaborated on in Chapter 7. In the case of failure of signalling equipment, permissions and orders can be exchanged in the form of written orders (which are nowadays generally dictated via a telephone connection). A train that could not be given a route will encounter a signal showing a warning aspect (yellow), later followed by a signal with a stop aspect (red). The train has to wait until the route can be cleared or it has to be given another route by the traffic controller.

Automatic train protection
Noticing signals and responding to them is clearly a driver's task, which is also safety-critical to avoid that two trains are on the same track section. As driver error can never be completely ruled out, this task was a single point of failure, until most railways developed a system to transmit the signalling information to the train in a direct way. These provide a parallel information stream which goes directly from the infrastructure to the train, so that the train can perform an emergency stop in those cases where a driver does not react properly to a signal. The railway's way to cope with driver error traditionally was to build in an automatic train protection system, thus bypassing the driver *and* his errors for safety-critical tasks. The most intensively used part of the Dutch network is equipped with ATB-EG (First Generation Automatic Train Protection). This system will be further explained in section 9.4.1. Its function basically is to check that a driver reacts to a yellow signal by applying the brakes. An often unexpected peculiarity is that the system cannot detect whether a train runs through a red signal and it will thus not intervene (further) if this happens.

Two other systems that are in use in the Netherlands are ATB-NG (New Generation Automatic Train Protection) (Poort 1992) and the European Train Control System (ETCS; Van den Top and Sierts 2009b). Both systems provide cab signalling, thus making trackside signalling superfluous, although this advantage is not yet widely made use of in practice. Also they provide braking curve supervision based on train's braking performance, which is an advantage over lineside signalling which has to assume the train with the worst possible braking characteristics. ETCS also provides a standardised interface to the driver and to trackside equipment. The two ATB versions and ETCS are technically incompatible, but a train with one of the newer systems installed can normally also use lines which are equipped with one of the older systems. To understand this thesis, no deep knowledge of these train protection systems is required, although it is useful to know that ETCS is the standard system for signalling and train protection throughout Europe which will be deployed in the coming decennia.

It might seem straightforward to simply install ETCS so that the SPAD problem is solved. In theory, this is a solution. In practice however, this is an expensive and therefore tedious process: there has been a call for nationwide implementation of similar systems since 1992, when a railway accident in Eindhoven happened, but not much progress has been made since it is considered a too expensive solution. In its rail safety policy document, the Ministry of Transport indicates that *'choices with regard to the replacement of ATB will mostly be based on capacity management, interoperability, controllability of the operational processes, information about the actual service provision, replacement of existing systems, running at speeds higher than 160 km/h and application of European state-of-the-art systems'* (Ministerie van Verkeer en Waterstaat 2004b). Thus, ATB-NG and ETCS are currently only

installed on a few lines, but these are not the lines where the problems requiring better utilisation are most urgent; also, because of their braking curve supervision, there are hardly any SPADs at places where these systems are installed. Although this of course is quite an achievement, it will not help us to understand how SPADs occur at those places where it is not installed.

Also, this thesis will show that installing such systems is only half of the answer: firstly, it does stop harmful consequences of errors, but it does not lead to an improved quality of service. Secondly, it does not take the error source away. Analysing the source of errors will lead us to many underlying weaknesses in the railway's organisation, which need to be cured not only to stop SPADs from occurring, but also to keep the system running at all in the future.

Vital and non-vital systems
In railways, a clear distinction is generally made between control systems and safety systems (Zigterman 1985). The process control systems are meant to use the available means of production as efficiently as possible to produce high quality transport services, whereas the safety systems serve as a backup so that errors in the control process do not lead to accidents. These systems are also referred to as, respectively, non-vital and vital systems. They are at the same time clearly distinguished and highly interrelated: vital systems give a boundary to what can be done by the non-vital systems, whereas the non-vital systems are used to steer the system towards a desired state within that boundary. Because both systems are so interwoven, they are often phrased together, for instance in German as 'Leit- und Sicherungstechnik' or in Dutch 'Beheersings- en beveiligingssystemen'. Often it is argued that, as vital systems safeguard the system's safety, changes in non-vital systems need not be checked for their influence on the safety level.

However, research in road traffic indicated that the safety effects of new measures cannot be analysed by looking at the new measures alone; they must be assessed from a broader scope taking also the needs and processes of the drivers into account (Jagtman 2004). Also one of the main conclusions of this thesis is that safety can be affected through changes in (non-vital) process control both in a negative and a positive way.

2.3.9 Discussion on the cascade model

A first note is that the cascade model is not fixed. Some functions can be performed in a different order or by different actors. For example, the Swiss railways design their timetable before they design their infrastructure. Another, Dutch, example where functions are performed by different actors is on some rural lines, where drivers set routes themselves. The function's order and actors can be changed, but the functions themselves remain more or less the same.

Further, it is remarkable that this description of the railway sector has demonstrated that there are hardly any provisions for feedback about the quality of the chosen control measures, as is indicated with dashed grey lines in Figure 2.4. If this is the case, the higher level actors will not know whether the solution they have chosen to solve a certain problem did indeed help the actors at the lower levels and so has led to the desired result. The main focus of this thesis will therefore be at the lower end of the cascade model: how well are drivers and traffic controllers able to perform their task in slightly disrupted operations? How do the outcomes of earlier processes in the cascade help them with that? The cascade model will be used to structure the different actors and functions throughout the remainder of this thesis. Indeed, one of the main findings of this research is that feedback from the bottom of the cascade

towards the higher levels must be organised. Since the focus of this research is on the operational level (i.e. the bottom-right side of the cascade model) the model was not fully developed on the top-left side. Several reviewers indicated that the level of system design is in fact not the highest level of abstraction, but that this process is also subject to higher-level restrictions, such as available finances or political decisions. This is true and during the course of this research project, several other functions became clear, so that an updated version of the cascade model will be presented in Chapter 10.

2.3.10 Related railway research

Much research is performed and publications are available to improve railway operations, but each tends to focus on only one of the processes in the cascade diagram. Examples are:

- analyses of signalling systems (IRSE 1995; Theeg and Vincze 2007);
- how long-term infrastructure capacity can be defined (Huisman et al. 2002);
- how timetable design can be optimised (Goverde 2005; Vromans 2005; Hansen and Pachl 2008);
- how traffic management systems can be developed, with a focus on mathematical algorithms (Van den Boom and De Schutter 2007; d'Ariano 2008) or on implications for traffic controllers (Lenior 1993; Neerincx and Griffioen 1996; Kauppi et al. 2006; Lenior et al. 2006);
- how drivers can remain alert, how they watch signals and define their driving strategy (van der Flier and Schoonman 1988; Banverket 2002; Human Engineering 2003; Metzger and Vorderegger 2004; Elliott, Garner et al. 2005; Hamilton and Clarke 2005; Li, Hamilton et al. 2006).

Also, much effort has recently been put into rail human factors research, especially in the United Kingdom. Compilations of many useful contributions can be found in Wilson et.al. (2005; 2007b). More specific references will be made throughout this thesis.

Of course, all of the studies mentioned provide useful information; nevertheless, research into the functioning of the cascade as a whole is the exception rather than the rule. Examples that cover more than one of the functions can however be given, such as:

- research performed by the Swiss railways and ETH Zürich, which starts by defining a service intention to customers and goes on to integrated timetable design, traffic management and train control (Lüthi et al. 2006; Laube et al. 2007);
- human factors research, aimed at understanding co-operation of traffic controllers with other actors (Farrington-Darby et al. 2006; Roth et al. 2006; Farrington-Darby and Wilson 2009);
- Research carried out by ProRail's traffic control department, taking into account timetable design norms and their application, route setting and driver behaviour (Schaafsma and Weeda 2009);
- Arco Sierts' contribution to ProRail's contest on how to make rural lines more attractive (Sierts 2007);
- the multi-SPAD analyses by the sector wide SPAD workgroup (see section 2.2.7) (Carstens et al. 2007);
- SPAD incident investigations (Dabekaussen 2005; Onderzoeksraad voor de Veiligheid 2005a; Human Engineering 2006; Van den Top and Groeneweg 2007; De Raadt 2008).

These publications will be cited at appropriate places in the remainder of this thesis.

2.4 Conclusion

Organisations in the railway sector
This chapter has described how the rail sector is organised. In summary, some organisations in the railway industry are, 15 years after its reorganisation, still searching for their raison d'être. Both main players, ProRail and NS Reizigers are still searching for their role and how to measure their performance. Also the relationship with the Ministry of Transport is still changing. With regard to the wishes of a better track utilisation by the existing system, it appears necessary to have clearer ideas of what 'good utilisation is': is it a higher customer satisfaction, higher profits, a higher market share or a higher cost/benefit ratio? The goal is not to maintain a railway network or to operate trains, but to create order in the time/space domain in which attractive transport service for customers can be supplied. 'Better utilisation' clearly has wider dimensions than just operating a higher number of trains on the same network.

Key performance indicators
No key performance indicators are set on the railway system (or even public transport system) as a whole; for example point-to-point travel time or point-to-point travel time reliability. Since these are indicators that are highly relevant to service perception by customers, it appears to be a good idea to introduce such key performance indicators (KPIs). Although there is a tendency to assess performance more and more in terms of customer satisfaction and/or in indicators that relate to the actual service output to the customer, this tendency seems to be merely used as a new way to measure performance in hindsight, but the customer's processes that need to be supported through the services offered, are not yet the starting point for what these services should actually provide and how they should be produced. The current KPIs do not yet have the right level of abstraction; neither do they exactly represent to what extent the customer need is fulfilled, nor do they give operators a clear grip on how to do their work. In such a way, the KPIs can become a goal in themselves, especially since the KPIs are the indicators on which the railway sector's performance is assessed by the Ministry. The result is that one only gets (nearly) exactly what the KPI prescribes. Safety performance appears currently to be assessed more or less in isolation from the other performance indicators, both by governmental regulations and by the sector's KPIs, which brings the risk that an improvement of one KPI comes at the cost of another.

Production of services
Railways are a synchronised system. The timetable, which makes the synchronisation explicit, therefore plays a central role in both the transport services market and in the path market. In both markets, customers and producers of the services should be able to produce their services simultaneously and in a coordinated way in order for the final service, transportation for the traveller or moving goods from A to B, to be satisfactory and safe. Within the railway's production processes, there is a strong top-down control, which is dominated by norms, standards and plans. This raises the question whether real time operators, such as traffic controllers and drivers, have clear goals and sufficient support to perform their tasks. Is it clear to them what they should do in order to meet the top-level KPIs and to deliver good quality of service to customers? There are hardly any indications of feedback between the different functions that were identified in the cascade model. If there is insufficient feedback, it cannot be guaranteed that the cooperation between the several functions works fluently, since higher-level actors do not know about problems on lower levels. Lacking feedback may therefore result in both safety and quality problems and a lack of learning.

Connection to the remainder of the thesis
Given the fact that problems in any of the levels of the cascade diagram will show up during operations and the fact that there appears to be little feedback of such problems, this will be the starting point for further analysis. The focus of the remainder of this thesis will therefore be to assess the cooperation of the several functions in the cascade model, for example to:
- understand which information operational staff need in order to steer the system in a safe way (top-down in cascade model);
- make management and designers understand how risks are dealt with at the operational level (bottom-up in cascade model);
- understand SPAD scenarios from a perspective of dynamic system behaviour.

In order to create this understanding, a new risk control model will be developed in Chapters 3-6; Chapters 7-9 will show how operations currently work, which problems operators face and how these are dealt with on various levels. Chapter 10 provides proposals for improvements.

3 Analysis of existing risk control models

3.1 Introduction

In order to understand how safety risks arise and how quality of service is assured, a model is needed in which the main features of both risk and the delivered services can be incorporated. Because this research intends to assess the influence of real time process control on these issues, this must also be incorporated in a suitable model.

This chapter starts with basic concepts of safety science, from which requirements will be formulated for an accident model that is suitable to answer the research question in this thesis (section 3.2). A review of existing approaches to understand accidents will be given in section 3.3. Early safety models considered accidents as a chain of events that eventually lead to an accident. Starting from the 1980s, accidents were more and more seen as inevitable by-products of the industrialised world: safety was not regarded as a goal in itself, but as a condition that should not be breached when one is doing one's business. Latest developments in safety models show that models are needed that can cope better with dynamic behaviour of the systems that are being modelled; a recent notion is that accidents are a result of failing control. Both notions, that of dynamic behaviour and that of (failing) control, fit well to our purpose of investigating how better process control can improve safety and performance. This chapter will end with the conclusion that a systemic model is required to incorporate all requirements.

Therefore, such a model, the 'safe envelope of operations' will be further developed in Chapter 4, and Chapter 5 will then focus on how to control the process within the requirements following from the safe envelope model. This will give a framework to understand how the railways cope with dynamic events.

3.2 Basics of safety science

3.2.1 Risk

Safety is generally defined as the absence of risk (MIL-STD-882D 2000; Reason 2000). In order to improve safety, it is necessary to have a good understanding of the contributing

factors that underlie risk. Kaplan and Garrick (1981) define risk as a set of *N* triplets, each consisting of:

- a scenario s (what can happen);
- a probability p that indicates how likely it is that a particular scenario will happen; and
- the consequences x that arise if it happens.

To allow for incomplete knowledge, they also add a scenario *N+1,* which represents all unknown scenarios. These can then also be attributed a probability and a consequence.

$$R = \left\{ \left\langle s_i , p_i , x_i \right\rangle \right\} \qquad i = 1,2,\ldots,N+1 \qquad \text{(formula 1)}$$

Others define risk as a multiplication of probability, consequences and exposure (see e.g. Henley and Kumamoto 1981; OECD 1997). This results in the following formula:

$$risk\left[\frac{consequence}{unit\ of\ time}\right] = exposure\left[\frac{events}{unit\ of\ time}\right] probability\left[\frac{adverse\ outcome}{events}\right] magnitude\left[\frac{consequence}{adverse\ outcome}\right] \qquad \text{(formula 2)}$$

Although risk is commonly considered as a multiplication of the factors mentioned before, Kaplan and Garrick are reluctant to multiply probabilities and consequences, as this reduces risk to an expected value of loss; this measure does not contain any information about the variance of the expected loss (e.g. many small injuries or a few people dying) and is therefore too simple.

3.2.2 Probability

Probability is a measure for the uncertainty of our knowledge. It says, for instance, that on average every 1 in *n*-thousand signal approaches will result in a SPAD. This figure depends on a probably large number of not completely understood mechanisms that are active in the system under consideration. The probability should not be taken as a fixed value as if it were a throw of a dice. It will vary, depending on the situation, especially if human operators are involved. When certain properties are already known (e.g. that the train driver is tired and that the signal is obscured by a tree) one should change the failure probability accordingly, for instance using Bayes' theorem (Bayes 1763). Probability will then remain a 'clean' measure for our uncertainty.

Confusingly, sometimes a probability per unit of time is used, which is an implicit multiplication of the probability and the exposure. The examples below make the difference clear:

- probability per approach: when 1 in 10,000 signal approaches results in a SPAD, the SPAD probability equals 0.01% per approach. This measure is useful when trying to understand the real-time process.
- probability per year: when there are 300 SPADs per year and there are 7,000 drivers, the probability for a driver to have a SPAD equals 300/7,000 = 4.3% per year. This measure is useful for off-line assessments on a more abstract level.

The 'short-cut' of a probability per year does not require detailed understanding of the processes in which the risks occur. This can be an advantage or a disadvantage depending on the purpose of the calculation. Since the purpose of this thesis is to create understanding, so as to be able to change the level of control, we must understand what is happening inside the black box. In this thesis, the formulation of probability per approach will be used. It will then be possible to find additional mechanisms of how adverse outcomes arise, so that targeted improvements can be made.

3.2.3 Exposure

Risk is often, but not always, defined per unit of time. In that case, exposure must be added to the triplet: a measure that indicates how often a process is performed within a certain time span, or, in other words, the number of opportunities for failure per unit of time. For instance, the SPAD risk per unit of time clearly increases when the number of train runs per year is increased; the same holds when the number of red signals per train run is increased, assuming that the failure probability per signal approach remains the same. Even if accident probabilities per event are low, absolute accident numbers can still be high if the exposure is high (DeBlois 1926).

3.2.4 Consequences

Accidents are a result of undesired releases of energy, which then can cause damage. This was already made clear by DeBlois in 1926 and was further explored in the 1970's (Haddon 1973; Johnson 1973). According to Johnson, damage can occur when one of the following types of energy are released: kinetic, chemical, thermal, electrical, ionising and non-ionising radiation, acoustic, biological; and also when interference occurs with normal energy exchange, e.g. in the case of exclusion of oxygen. Also the Accident Deviation Model (Hale and Glendon 1987), which can be used to describe consecutive accident phases, considers uncontrolled release and transfer of energy as the core element of accidents.

The energy that can be released and cause damage can be the same energy that is necessary to keep the desired process running (e.g. in transportation) but it can also be energy that is present in the system or in the environment for other reasons (e.g. when someone gets struck by lightning). Transport, by its very nature, involves passengers carrying kinetic energy: they even pay for it. Since the energy cannot be taken away transport accidents can always have severe consequences. Although measures can be taken such as crashworthiness of vehicles and protection of passenger in the vehicles, this is outside the scope of this research which focuses on reduction of accident probabilities and not on its consequences.

3.2.5 Acceptability of risk

Although it is often said that safety comes first, this is demonstrably not how the world works. People balance risks and the benefits that come with an activity. This was observed by Wilde, who described the phenomenon in his risk homeostasis theory (Wilde 1994), which is also sometimes referred to as risk compensation theory. Rasmussen (1997) noticed similar behaviour for organisations, which constantly migrate towards the boundary of acceptable performance, due to market forces. Kaplan and Garrick (1981) argue that acceptability of risk cannot even be considered in isolation; it must be traded off against the expected benefits of the activity and the options that are available to decrease risk, because no risk is acceptable if it can be easily prevented and/or if there is no benefit attached to it. An important notion here is that the observed trade-off behaviour is guided by the operator's or organisation's belief that they are still in control of the situation. This supports the thesis that accidents can be considered as a failure in the control system, as was stated by Rasmussen (1997) and Leveson (2004).

3.2.6 Risk management

There is a fundamental problem with defining safety as the absence of risk or the absence of accidents. The safer a system gets, the less information is available about how safe it actually is (Reason 2000). It was therefore proposed by Reason and later also by Hollnagel et al. (2008) to define safety not as the absence of risks or accidents, but as a positive measure, e.g. 'the ability of individuals or organisations to deal with risks and hazards so as to avoid damage or losses and yet still achieve their goals' (Reason 2000). In other words, safety is the

ability of individuals or organisations to control their processes and thus decrease the probability of failure. This is well in line with the notion that accidents can be attributed to a failure in the control system.

3.2.7 Requirements for a risk control model

To understand how risks arise in a dynamic system like railways, a model should meet the following requirements.

1. The model should describe the operational level with all its relevant variables, in which operators are controlling the events in the system. Given the design and organisation of the system, these are the people who are the last to balance the risks and the benefits of the activity. They can decide how the activity will progress.
2. The model must be able to represent quality of service (the goal that the operator should meet).
3. The model must be able to represent the energy contained in the system and in its environment (in order to model accidents).
4. The model must be able to reflect earlier choices, such as those at the design and management level. These define and influence the conditions in which operators have to do their work, such as operating procedures, technical aids and the means that they have at their disposal to control system processes.
5. The model must be able to respect relevant influences from outside the defined system boundary.

3.3 Overview of existing approaches to explain accidents

3.3.1 Several classes of approaches

The previous section has made clear the basic concepts about risk and its underlying factors. Yet we still have not seen models or approaches that explain how these risks arise; this section will describe existing approaches. Accidents were long time considered as a matter of chance or an act of God. They just happened every now and then. This is a rather passive view in which no causal relations are considered (or they are ascribed to divine intervention) and the consequences of accidents are more or less accepted; hence this ‘explanation’ gives no clear starting point for how safety can be improved. This approach is therefore not useful. During the past decennia, there were several paradigms and models that have tried to explain accidents, which will be mentioned briefly. For a more detailed overview, refer to Aven (1992), Kjellén (2000), Rasmussen (1997), or Van Alphen, Gort et. al. (2008).

During the years, many accident models have been developed, so that it is impossible to describe them all separately. Also, this section will not discuss what the definition of an accident model should be in the abstract; this thesis sticks to the notion that a model is a simplified abstract view of the complex reality. Which factors are relevant in a model and which depth is required for the model, depends on the purpose of the model and can therefore always be discussed. Distinct classes of models can be distinguished, such as whether the model can be used in a prospective or retrospective way or whether a model’s main focus is human, technological or organisational failure. However, for the purpose of this thesis, this is not the essential difference between most models, since we want to use a model to gain understanding (for prospective or retrospective analyses) and control in a socio-technical system, which involves human, technology *and* organisations. A useful classification is provided by Hollnagel (2004), who distinguishes between sequential, epidemiological and systemic approaches; this classification will also be used in the next sections as the framework for discussion. Finally, also some basic but fundamental knowledge about human

error will be provided, since this is important knowledge to interpret the remainder of the thesis.

3.3.2 The sequential approach

During the industrial revolution, when new technology was rapidly introduced, it became clear that there must be a cause for an accident. Two views on safety became common: one of them focused on improvements in technology and one on human error. Many technical improvements to machinery were made, so that frequent accidents became less likely or even impossible when an inherently safe design was chosen. Safety devices were installed, such as barriers to protect workers from moving machinery parts. Many of today's basic safety concepts in railway operation have their origin in this period, such as the interlocking principles or the automatic pneumatic brake (Jongerius 1993).

More or less simultaneously, the human error view emphasised that workers had to be alert and should be disciplined. This view culminated in the accident-proneness-theory, which stated that some people have characteristics which make them more prone to accidents. This idea does not lead to many proposals for improvement of safety, since the only way to get around the problem is to ban such people from the process. The theory was abandoned after severe criticisms, mainly that only one factor to explain accidents is chosen ex-ante, and that some people are involved in more accidents than others by pure chance (Swuste et al. 2008). A related viewpoint regarding human behaviour was the idea that accidents occurred because people violated rules. Possible improvements then were the introduction of more, or different rules and training procedures: the people were changed to fit the system. For instance, in the case of a SPAD, it is often said that the driver failed to notice the signal. However, in fact it is *arbitrary* whether this is perceived as the driver failing to see the signal or the signal failing to attract the driver's attention (Davis 1966) and thus underlying causes and solutions may also be found in the design of the signalling system. Asking further questions usually leads to other causes being identified (Dabekaussen 2005; Kletz 2006).

In the early 20th century, it became clear that the sequence of events leading to an accident is generally longer than just the last act before the accident or incident happened. Also factors from the environment, design and management may have their influences, as was illustrated by Heinrich's domino model (Heinrich 1941). In the Hazard – Barrier – Target model (Haddon 1973), energy can flow from a hazard to a target. Since we do not want this to happen, a barrier must be placed in between. It is easy to think that a barrier must be a physical object, such as a fence or personal protective equipment, but according to Hale et. al. (2007) this is not the case, since anything that prevents the hazard from reaching the target can be considered a barrier; for example skilled behaviour to keep out of the vicinity of the hazardous energy. Nonetheless, the barrier must have a physical consequence in some form or another to prevent the transfer of energy.

Probably the most well-known metaphor to illustrate this is the frequently used Swiss Cheese Model (Reason 1990). This model defines several layers of defence to prevent accidents, ranging from organisation management, via workplace factors to personal factors. Because

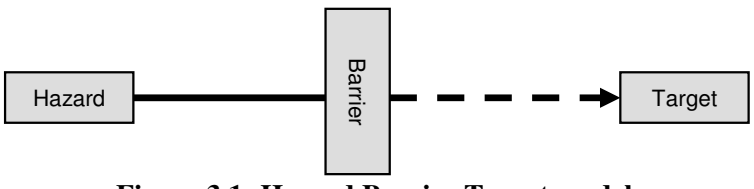


Figure 3.1: Hazard Barrier Target model.

none of these layers are perfect, there are holes in them (hence the name of the model) and if the holes line up, accidents can happen. This resembles the hazard – barrier – target model. Such holes are also called latent conditions: these are properties of a system that do not directly lead to an accident, but they make it more likely to happen. They are present well before an accident sequence becomes visible. Examples are a poor safety culture or design errors in a human-machine interface, maintenance failures et cetera. Latent conditions may render different kinds of protective measures ineffective so that an accident can then happen. From the Swiss Cheese Model it follows that often many layers will have failed before an accident could happen. Although it may then be concluded that there were many shortcomings in a system, a more positive view is that it is often sufficient to repair just one barrier in order to prevent similar accidents from happening again. MORT (Management Oversight and Risk Tree; Johnson 1973) advances on the idea of energy flows that get things done in a system (Kingston et al. 2004). It therefore does not only focus on what could go wrong, but also on what people wish to achieve with the process: safety and performance are two facets of the same management process. MORT clearly tries to integrate several levels, such as design, management and operations. It comes up with eleven basic possibilities to avoid damage (Johnson 1973):

1. limit the energy (or substitute a safer form)
2. prevent the build up of energy
3. prevent the release
4. provide for slow release
5. channel the release away – that is, separate in time and space
6. put a barrier on the energy source
7. put a barrier between the energy source and men or objects
8. put a barrier on the man or object – block or attenuate the energy
9. raise the injury or damage threshold
10. treat or repair
11. rehabilitate

Not only did such developments lead to the insights that accident sequences can be quite long; another insight was that in some cases an event leading to an accident may only happen if two or more conditions are fulfilled. Such multiple causes could be modelled as fault trees; similarly, multiple consequences were modelled as event trees. These are the basis for several widely used methods such as the bow tie model (Visser 1998) and TRIPOD-Beta (Groeneweg 2002). The basic concept of barriers has also been transferred to, for example, the bow tie model (Hale et al. 2007).

A fault tree specifies which (failure) conditions must be met for the undesired event, the so-called top event, to happen. If the path to that event is blocked at some point, the accident is prevented. In the fault tree handbook, also a success tree is mentioned. It is ‘a logical complement of the fault tree, with the top event of the success tree being the complement of the top event of the fault tree’ (Stamatelatos and Vesely 2002). A railway example of such a top event in a fault tree would be a SPAD; in a success tree, the top event would then be ‘no SPAD’. However, this is only a success from a safety viewpoint: railway systems were built for another purpose than ‘not SPADing’, such as transporting passengers or goods safely from A to B in a commercially attractive way. A success therefore has to be defined in terms of the system’s purpose. Safety is only an (important) boundary condition. Also an operator of a system wants to know what he must do in order to keep the system safe *and* running.

Event trees often are considered to be the right-hand side of a bow tie, so that they start with the undesired event and show which adverse consequence may result. However, they can also

start with any system state and will in that case show which possible system states (wanted or unwanted) may result and attach a probability to each branch. Although this can give a good overview of possible outcomes, it relies on (given) failure probabilities. As such it does not explain which mechanisms underlie these probabilities. As a consequence, an event tree is neither predictive nor explaining.

Tree-shaped models have limitations when coping with dynamic system behaviour, since they assume a fixed order in which events take place; if events occur in a different order, the tree would have a different shape and the probabilities contained in it may change as well. Although they can represent different orders, they do not help to understand which order is worthwhile to take into consideration. Also, they often assume component or operator failure as the decisive criterion to choose between the 'right' and the 'wrong' branch of the event tree, and thus put a relatively strong emphasis on this type of failures. These models do not take into account what incentives there are for an operator to choose for one branch or the other. Finally, tree-shaped models do not easily cater for recovery of errors (Schwarz et al. 2009). Although such events could be added, these make the tree very complex.

A sub-category of sequential models form the deviation models. These focus on several phases of which accident processes consist. Examples are the Accident Deviation Model (Hale and Glendon 1987) or the OARU (Occupational Accident Research Unit)-model (Kjellén 2000). Also quality management systems can be considered as deviation models, since they prescribe a certain process from which deviations have to be detected and solved, in order to assure quality.

In descriptions of normal operations, incidents and accidents are often considered as a by-product. The opposite is the case with sequential accident models: these generally focus on what went wrong and do not consider normal operations. Sequential models are, due to their linearity, limited in their capability to explain complex socio-technical systems in which all kind of interactions occur. Rasmussen (1997) and Leveson (2004) argue that event chain methods generally do not account for:

- social and organisational factors in accidents;
- system accidents: accidents that usually do not result from component failure but from unintended interaction between sub-systems;
- human error and adaptation: people changing their behaviour as a result of safety measures which causes the measures to be less effective (a fact represented in the risk homeostasis theory).

3.3.3 The epidemiological approach

Epidemiology is the scientific study of the occurrence and spread of diseases among the population. The discipline tries to find out who is affected by a disease and what factors play a role, (age, sex, infection sources, food, etc.). The English Dr. John Snow was famous for discovering a cause of a cholera epidemic in the Broad Street in London's Soho in 1854. By questioning cholera patients and by drawing cases on a map, he concluded that the source of the infection outbreak must have been the local water pump. He made an end to the epidemic by removing the handle of the pump (Snow 1854). Note that Dr. Snow did not discover the causative agent of cholera, a bacterium that was discovered later.

Epidemiological accident research tries to explain accidents in a similar way: injured are those who are exposed to causal factors in their environment. This approach makes use of models such as Haddon's Hazard-Barrier-Target model. The epidemiological approach to understanding accidents is to look at the influence of possible contributing factors: therefore,

it is necessary to collect data on many possible factors and hope that a relationship between the factors and the problem under consideration can be found. Statistical analyses of large databases are often an important tool (Saari 1986). Therefore the epidemiological approach only allows finding relationships between accidents and possible contributing factors on those items that have been used for data collection.

There are also several methodological problems with the epidemiological approach. The required large datasets are not always available in safety science, especially not in sectors that already perform relatively well (Saari 1986). However, if there is a relationship between incidents of a certain type and accidents, the dataset can be filled with incidents, too, which alleviates the problem to a certain extent. Also, the exposure and the amount of non-accidents are normally not registered. An epidemiological approach might for example find that most pedestrians are killed on zebra crossings, so that something must be wrong with them – not taking into account that most successful road crossings are also likely to take place on zebra crossings. Similarly, statistical analyses are also susceptible to reversed causation (e.g. having gray hair is healthy because people with gray hair grow old) or common causes (e.g. people having grey hair are likely to suffer from Alzheimer's disease); these drawbacks can easily occur when taking the epidemiological approach without any causal mechanism that is being looked for. The epidemiological approach does not necessarily lead to deep causal explanations and increasing safety appears to be possible mainly through reduced exposure to the alleged causal factors. For example, the cholera outbreak could be stopped by removing the handle from the pump. Although this solved the problem, it did not find the real cause of the illness. Hence, the approach can be helpful to provide a good starting point for further investigations: for instance to explore what makes the water from this pump different than that from others.

In summary, two main drawbacks of the epidemiological approach to explaining accidents are the necessity of large datasets and the fact that exposure can not easily be taken into account due to the non-availability of non-accident data. An advantage is that if large datasets are available with sufficient possible influencing factors, they can be used to see where to look further.

3.3.4 The systemic approach

Although both sequential and epidemiological models have led to useful insights and can be used for certain types of problems, they do not lead to a full explanation of why accidents happen. James Reason, the 'inventor' of the Swiss Cheese Model, indicates that *"predictive accident models must (…) not only recognise that confluences occur but also provide a plausible explanation of why they happen. (…) The inescapable conclusion is that we need accident analysis methods that look equally to individual as to organisational influences. In other words, models of 'human error' and organisational failures must be complemented by something that could be called socio-technical or systemic accident models"* (Reason et al. 2005). Rasmussen showed that productivity requirements push organisations towards the boundary of safe operation and that therefore new, dynamic models are necessary to describe safety in its (changing) context; these models should focus not only on what goes wrong but also what should be achieved. They should be based on control theory (Rasmussen 1997). Systemic accident models are therefore not based on barriers or linear causality, but on incorrectly performed functions or unintended interactions between functions. These functions can be performed by humans, machines or organisations.

Where sequential models assume a fixed order in which events take place, this order may in fact be subject to changes depending on the order in which functions are fulfilled and to what

extent. For example, a train driver can decide to brake well ahead of a signal, which allows him to double-check the signal aspect and correct his behaviour if necessary. He can also decide to start braking only after passing the signal. This would result in two different event sequences, but in fact, the same functions are performed. Similarly, the driver could decide to brake hard or just slightly, which is difficult to depict in sequential models. Therefore, we can describe sequential models as static whereas systemic models are better capable to cope with dynamics in the system. Such dynamics can be important. Perrow (1984) indicated that in complex (as opposed to linear) systems, failures propagate in unpredictable ways. In systems that are tightly coupled (as opposed to loosely coupled), such errors also propagate almost immediately to other subsystems. This makes the behaviour of tightly coupled, complex system incomprehensible, and hence accidents in such systems are normal. However, other researchers (Weick et. al. 2002; de Bruijne et al. 2006) have found examples of complex and tightly coupled systems, such as nuclear power plants and aircraft carriers, that have a good safety record. These so-called High Reliability Organizations or High Performance Organizations share certain properties which make them resilient to failure. Although High Reliability Theory and Normal Accident Theory are often considered to be opposites, this need not be the case. We should, instead, consider that complex and tightly coupled systems need a special control structure, which enables them to handle unforeseen situations. Apparently, such control structures were not (yet) functioning well in the systems that were analysed by Perrow.

Several systemic models, such as Acci-Map (Rasmussen 1997) and STAMP (Leveson 2004) have been developed. However, Acci-Map is not dynamic and thus not well able to show which control options operators have. STAMP takes a new way of looking at things, since it does not focus on unsafe events but on constraints that should be imposed on system operation: ‘unsafe behaviour results from either a missing or inadequate constraint on the process at the lower level, or inadequate enforcement of the constraint leading to its violation’ (Leveson 2002). It is a recursive model as any process will again have limitations from a higher-level process, such as design or legislation. Finally, the safe envelope of operations (Hale et al. 2007) is a new, emerging model. It describes the system as a state space in which several areas are defined to explain accidents. The concept tries to incorporate design and management influences and system operation, by combining advantages of several models that we have mentioned before. This is the model which will be worked out further in the next chapters.

3.3.5 Human error models

Starting in the 1970s, cognitive psychological research pointed out how humans mentally process information and make decisions (Hale and Hale 1970; Hale and Glendon 1987). This provided a basis for better understanding of human error and led to some human error classification schemes. Well-known are Rasmussen’s SRK-model (Rasmussen 1983) and Reason’s Generic Error Modelling System (GEMS) (Reason 1990). The SRK-model (skill, rule, knowledge) says that humans first try to perform a task on an automated, skill-based level. If this is not possible, a mental rule is used to decide on a proper reaction. If this is also not possible, an analogy with another system must be made or a more fundamental knowledge-based search must be initiated until an effective solution has been found. After a comparison of several studies into human error quantification, Hinzen (1993) concluded that the error probability is on average typically 1:1000 for skill based activities, 1:100 for rule-based actions and 1:10 for knowledge-based ones. However, the real error probability has a large variance around this mean value and can vary from approximately 1:50.000 actions to almost 1:1, depending on many circumstances such as experience, time for error detection, clearness of information provided, etc. (Kirwan 1994).

Recent research in the area of neuropsychology appears to confirm the phenomena described in Rasmussen's (SRK) and Reason's (GEMS) human error models. There are different areas in the brain for routine actions, planning and memory (Klein 2006). This appears to correspond fairly well with the error types identified by Reason. If someone is asked to solve a complex question for the first time, some of the brain cells have to be reconnected into a new network, until a solution has been found. This reconnecting requires energy, which is catered for by an increased supply of blood sugar and oxygen. After a few repetitions, similar types of questions can be solved much easier and quicker, because the neural network has already been formed and less energy is necessary. In the case of motor routine actions such networks are not only formed, but also moved to different areas in the brain once sufficient practising has been undertaken. This makes the brain faster, more effective and cheaper in terms of metabolism. This also allows us to simultaneously perform such motor actions while the remaining brain areas are still available for more complex tasks (Roth 2006). In conclusion, consciousness is slow, error-prone and 'expensive' but it is the only way to solve tasks for the first time (Roth 2006). The brain always tries to automate sequences as much as possible and by doing so, to ban them from conscious behaviour. Although efficient, this also risks that automatic response is sometimes performed when it is not desired. Thus, correct structures in the brain can be achieved to a certain extent over time through training and experience, but the brain's functioning should by and large be considered in the short term as a given fact. Therefore it is of utmost importance that the system design supports the human operator. Human factors research therefore emphasises that accidents are a result of a mismatch between the human capabilities and the task requirements (Endsley 1999). From this viewpoint, it is often more fruitful to change the system to allow people to work in the best possible circumstances. This again indicates that human error, technical failure and design error are labels that basically (can) refer to the same thing: a function that is improperly fulfilled.

3.4 Conclusion

In summary, this chapter has provided basic concepts from safety science. Although accidents are clearly undesired, they should not be considered as something abnormal: risks are inherent to undertaking an activity and therefore it requires a continuous effort to prevent or control them, in which the process of undertaking the activity must be involved. Risks cannot be considered in isolation, but must be seen in the context of normal operations. It was shown that most existing models do not suffice to describe certain types of accidents in relation to normal system operation. They focus on failure of certain system elements, often in isolation (e.g. human operator, technology, design or management), but some accidents cannot be fully understood by doing so.

Table 3.1: Overview of existing types of accident approaches in relation to the requirements.

	Requirement	Seqnt.	Epidem.	Syst.	Human
1	Describe how operators control events and balance the risks and the benefits	No	No	Yes	Yes
2	Represent system's goal	No	No	Yes	No
3	Represent energy in system	Yes	Yes	Yes	No
4	Reflect design and management choices	Yes	No	Yes	Yes
5	Respect influences from outside the system boundary	No	Yes	Yes	Yes

In Table 3.1, the different types of accident models are compared to the requirements we need to help understand railway safety (e.g. SPAD incidents). We see that we need a systemic model. The literature on high-reliability organisations and normal accident theory focuses

mainly on organisational aspects, and therefore is not fully suitable to model the operational aspects that are important for the research question of this thesis. Acci-Map does take several control levels into account, but is a static model. Both STAMP and the safe envelope of operations, a new model in development, would appear to be suitable for the purpose of this thesis. Since the safe envelope model was proposed in Delft University of Technology where this research was undertaken, and the model required further development and testing to prove its value, it was chosen as the framework for the remainder of this thesis.

PART II: Theoretical Framework

4 An integrated risk control model: the safe envelope of operations

4.1 Introduction

The previous chapter contained a number of requirements for a dynamic safety model, which is necessary to fully understand how risks in a dynamic system like railways arise and how they can be controlled. It was argued that a model is necessary in which safety can be modelled in its context. After all, one usually does not want safety per se, but one wants to perform a certain activity in a safe way. This chapter presents such a model: the safe envelope of operations (Hale et al. 2007), a recent development in safety science which fulfils in principle the requirements mentioned in section 3.2.7. The chapter will deal with the following issues:

- The model's basic features will be explained, and it will be further clarified and detailed.
- An example will be given which serves both to illustrate the model and to improve it, by working it out more fully.
- The chapter will be concluded with advantages and disadvantages of the new model, compared to the requirements.

The model introduced in this chapter will be connected to other, existing control models in Chapters 5 and 6, and it will be used to analyse and integrate the findings from the practical part of this study in the Chapters 7 - 9.

4.2 Development of the safe envelope so far

A first representation of what is now called the safe envelope of operations can be traced back to research in a car driving task. Gibson and Crooks (1938) mentioned the 'field of safe travel' – a sort of tongue protruding from the car along the road as a safety envelope. This was a purely spatial representation.

In aviation engineering, the concept of the flight envelope is used to define the limiting operating conditions of a plane, depending on its installed power, its flight altitude, its air speed, and other variables. If the combination of all these variables constitute a point within

the flight envelope, this is considered a controllable and thus a safe situation (Walgemoed 2005). Another application of a similar concept during flight operations is the perspective tunnel display, which provides pilots with a virtual tunnel though which they should fly in order to follow the desired trajectory (Grunwald et al. 1981). The flight envelope in aviation is used in a mathematical way, in which the relationship between these variables can be calculated.

The concept of the safe envelope of operations that was developed more recently in safety science (Hale et al. 2007) can be considered as a further development of the notion of the flight envelope, and the Accident Deviation Model, combined with Rasmussen's notion that organisations continually migrate towards the boundary of acceptable performance: there is a competitive advantage if a company is able to continue working close to the boundaries where others have to stop. Whereas the flight envelope considers predefined variables and takes a technical view, the envelope in safety science is more generic and can also be seen from an organisational viewpoint, for example how well the organisation (or individuals within that organisation) is able to set the boundaries correctly and remain within them. Both approaches share the notion that a process should be kept within a boundary of controllable variable values in order to remain safe.

The safe envelope model as proposed by Hale et.al. (2007) represents the state of development of the model in 2005, before work began on this thesis. The published paper did not appear until two years later and did not include the developments made in this research. It modelled a system by defining three boundaries or envelopes to group certain system states:

- a safe envelope boundary, which comprises all system states in which no damage occurs;
- a defined operational boundary, within which the organisation or the person in control wishes to keep the activity;
- a boundary of controllability, indicating which system states can be effectively controlled by the operator.

This model appears to be useful for use in this research, although some issues have to be addressed to make this proposal fit for use. For example, several examples given in the 2007 paper actually contradict the definitions given in the same paper. An example is that uncertainty about where damage will occur is first being defined as uncertainty about the exact place of the safe envelope boundary, whilst later in case of uncertainty 'we designate a whole area as danger zone, even though not all of it is'.

In another case, examples are being given of controlling an airplane (i.e. keeping it within the safe envelope boundary) and it is stated that in case of a wind shear or an engine failure the safe envelope suddenly moves inwards, which means that there are suddenly more states in which damages occurs. However, in fact the safe and unsafe system states do not change themselves in such cases, since the states in which the plane gets damaged remain the same. However, it may be harder or even impossible to avoid these unsafe states, which says something about the *controllability* of the current situation and the *attainability* of future system states. Although one of the reasons to develop the safe envelope is to put more emphasis on real time control, the definitions of the concept and of the controllability need to be worked out in more detail, which will be done in this chapter and the next one.

Also another addition to the safe envelope could be made. From a pure safety viewpoint, it is sufficient to make sure that the process does not cross the safe envelope. However, from a 'production' viewpoint, it would also be interesting to provide the operator with information so that he can form an idea about where in that whole safe area he should be operating in

order to reach the goals of the activity. In other words, although not flying into the terrain is of prime importance, flying a route that leads towards the destination is also something that the pilot should take care of. As was defined in section 1.5.3, for this research a framework is required that can combine safety with 'production'. Therefore, this chapter will also add such elements to the safe envelope model. Finally, the safe envelope model in the Hale et.al. paper is presented in a generic way. It requires some adaptation and further detailed specification to make it suitable for use in a railway context.

4.3 Terms and definitions

The description of the concept will start with definitions of the basic terms. These will then be applied to defining and using boundaries and envelopes for operations, which will be made clear using an example drawn from railways.

System, system boundaries and environment
A system is a regularly interacting or interdependent group of items forming a unified whole. Generally, man-made systems serve a purpose. A system can be a technical system, but also an organisation, a person or a combination of these. A system is separated from its environment by system boundaries. When modelling systems, these are generally chosen such that only a limited interaction with the environment needs to be considered. The environment may consist also of other (sub)systems, e.g. if the system consists of one train, other trains are parts of the environment. For correct application of the safe envelope model, it is therefore important to define clearly which system is under consideration.

Essential variables
The essential variables are those variables that can be used to measure whether the system's goal and its safety have been achieved. These variables depend on the characteristics of the products or services that the system should produce and on the safety requirements. Note that the goal and the safety requirements are only met if the variable values are within a particular range, see below.

System state
The system state is a complete description of a system at a particular moment in time, in terms of its essential variables. For railways, this could for example be a description in terms of a train's position, its speed, and the restrictions of the route ahead. Which variables are relevant depends on the problem under consideration.

Process
A process is a series of actions or operations changing the system state (input) into another state (output). It thus affects the essential variables. Generally, an intentional process is performed by a certain (sub)system and leads, if executed correctly, to a desired system state.

Energy
System states cannot be changed for free: energy is required to keep the desired process(es) running. Energy also plays an important role in the (potential for) damage to the system and thus for its safety. The safe envelope model should therefore include all relevant types of energy that affect the system: these can be internal energies (e.g. to keep the process running) or energies in the system's environment that may affect the system (e.g wind, lightning).

Accident
Similarly as in MORT (Johnson 1973), an accident is defined as an event in which energy is *unintentionally* transferred from or to the system such that damage occurs in the system and/or its environment.

Boundaries, ranges and envelopes
For most variables there are certain *boundaries* beyond which the system cannot operate properly or will suffer damage. An example is the human body temperature, which should preferable lie between approximately 36 °C and 37 °C, but certainly not fall below 32 °C or exceed 42 °C. A *range* indicates all values of a single variable that are permissible. This depends on the objective. If the objective is to prevent the system from being damaged, this is the *viable range* (between 32 °C and 42 °C). If the objective is to keep the system running normally, we must keep it in the *operational range* (between approximately 36 °C and 37 °C). There can also be boundaries to a combination of variables, in which case the boundary of one variable is dependent on the momentary value of another variable. When we are not talking about a single variable, but about the system as a whole (with all its variables) we use the term *envelope* instead of range; see Figure 4.1. Each variable has a separate axis, so that there are generally more dimensions than can be drawn in two-dimensional paper images, which are therefore always a simplification of the model.

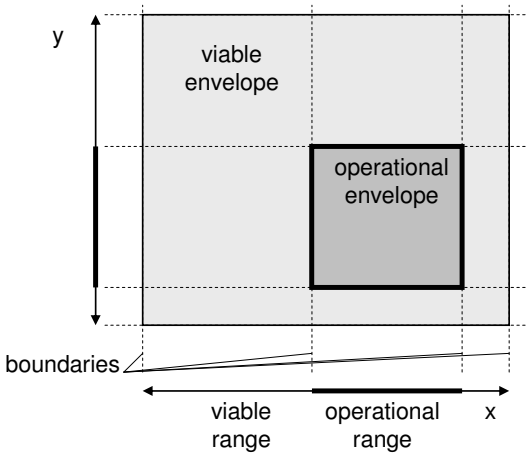


Figure 4.1: Ranges, boundaries and envelopes for a a two-variable system.

Control
All kinds of internal and external disturbances constantly change the system state, which might therefore go outside its desired envelope. Some form of control is necessary to prevent this. Process control is defined as *'the process of causing a system variable to conform to some desired value, called a reference value'* (ISO8402 1986). Franklin et al. (2006) define that a system *'is controllable if there exists a (piecewise continuous) control signal u(t) that will take the state of the system from any initial state x_0 to any desired formal state x_f in a finite time interval'*. For correct application of the safe envelope model, it is therefore important to define clearly:
- who or what is or are the intended controller(s) of the system under consideration;
- which result should be achieved at which moment.

For more information about control, refer to Chapter 5.

Types of envelopes
The safe envelope model defines the following envelopes (1-4 in Figure 4.2):

A. The *viable envelope (VE)*, which encloses all system states in which no damage occurs; i.e. all states in which no more energy will be transferred from or to the system and its environment than these can handle without being unintentionally damaged. The viable envelope is therefore partly dependent on the environment and partly on the properties of the system itself.
B. The *controllability envelope (CE)*, which encloses all system states for which effective control measures exist (designed or informally evolved) for the actor(s) and objectives that are currently being modelled. The objective can be to keep the system in the viable envelope or within the operational envelope (OE).
C. The *operational envelope (OE)*, which encloses the system states defined as acceptable by the organisation(s) responsible for the activity. The boundaries of the OE are often laid down in laws, regulation, manuals, procedures etc.
   Since not all higher-level functions (e.g. timetable design and infrastructure design) use the same acceptability criteria (e.g. different variables and/or different boundaries per variable), there can exist different OEs.
D. The *attainability envelope* (AE), which indicates which system states can be reached within a given time increment $t + \Delta t$, given the actual state.

Since one cannot assume upfront that variables are independent of each other, they should not be drawn perpendicular to each other as was done in Figure 4.1. To illustrate this, Figure 4.2 also contains some limitations that are dependent on several variable values. All envelopes can change over time. This emphasises that we are dealing with a dynamic process.

Control measures may also exist to limit the amount of damage once the process has left the safe envelope and a damage process is occurring, but we do not consider them here since the focus of this research is on combining quality of service and accident prevention. Therefore, in this thesis, a control measure is always a measure that can prevent an accident, under the condition that it is executed correctly and in time, and the controllability envelope must always be within the viable envelope. Within the chosen stance that unwanted transfer of energy is unacceptable, the operational envelope should always be smaller than the viable envelope; and it should be completely controllable. However, this is not always the case in practice, as Figure 4.2 shows.

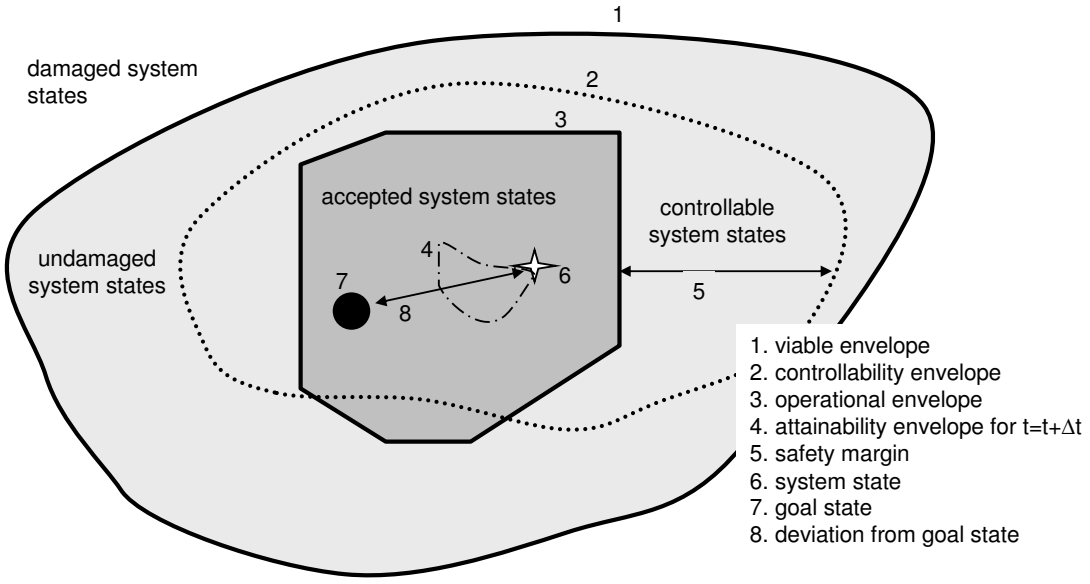


Figure 4.2: The model of the safe envelope, flattened to a two-dimensional image.

Safety margins
The safety margin (5) is the room between the controllability envelope and the operational envelope. A margin allows for uncertainty about the system itself or its control system (Starbuck and Milliken 1988). For instance, if the exact position of the safe envelope or the position of the process therein is unknown or if its controllability cannot be fully guaranteed, a margin may be applied to prevent small inaccuracies from leading immediately to damage. Safety margins are system states that are considered safe, but are deliberately not used under normal circumstances. Here, the trade-off to maximum production manifests itself, as this strives for operating under as many possible circumstances as possible, and thus minimising safety margins. In the states between the controllability envelope and the viable envelope, no accident has happened yet, but the intended controller of the system can no longer control the system state and it is thus not guaranteed that the system will remain within the viable envelope. Only chance events or intervention from outside the system being modelled can keep the system within that viable envelope and may be able to return it inside the operational envelope. These states are therefore not considered margins in the sense of the definition above.

Goal state and tolerances
Within the limitations indicated by the operational envelope(s), the system's operator is free to choose how he wants to control the process in order to reach the optimal system state, the *goal* (number 7 in Figure 4.2). Management usually decides which dimensions must be considered relevant to determine whether or not a state is optimal, e.g. with regard to multiple aspects such as safety, operational performance, punctuality and costs. Some of these are made manifest in an OE, thus indicating the allowed tolerance, some are left for the operator to decide on his own judgement.

Operators tend to optimise system behaviour in a heuristic way, based on various cues and experiences with similar situations, and under time pressure. They are therefore looking for a satisficing (near-optimal) solution and not necessarily for the optimum itself (von Meier 1999). To indicate whether a near-optimal solution is fair enough, the acceptable *tolerance* around the goal should be explicitly indicated. This can be modelled by means of an operational envelope. A well-known example of this is the 'estimated'-symbol (℮) on food packaging (e.g. 450 g ℮) which indicates a weight within a predefined range. The goal state and operational envelope can be used to model the required quality of service. The goal state and the acceptable tolerance may change dynamically, depending on changes in the actual situation and the criteria that are considered most important. The difference between the actual system state (number 6) and the goal state is the deviation from the goal state (number 8 in Figure 4.2).

Trajectories
Until now, the safe envelope of operations has been described for a single point in time, of which Figure 4.2 is an example. Sooner or later, all numbered items in Figure 4.2 can change. Therefore, the picture of the safe envelope is only a cross-section at a certain moment. Time can be represented on an axis perpendicular to the envelope. When time passes, the system's state will have a certain path through the envelope, as will the various boundaries. This is what we call a *trajectory.* Such a trajectory is the result of all changes that affect the system's initial state, regardless of whether this was intended by the operator or caused by other influences. Thus, it is possible to distinguish between an intended trajectory and an actual one.

4.4 Balancing attention between safety and quality

4.4.1 Quality and safety management

Quality management requires the trajectories to be steered towards the goal and within the operational boundaries. Safety management requires the trajectories not to cross the viable envelope. In both cases, this involves process control. Safety management can then be seen as a particular (and important) sub-task in quality management. Under the assumption that good quality implies safety, doing things right implies doing things safely. Although safety and quality of service can thus be treated similarly, it is not enough to focus on safety only or on quality only. An example of what may go wrong in the first case, is that there can be trajectories leading to system states that are neither desired nor unsafe, such as running no train services at all. A focus on safety only would not consider this as an unwanted state. In the other case, there may be trajectories that do lead to the desired quality, but which require the process to be controlled in such a way that it is susceptible to safety-critical errors and therefore unsafe (even if no accidents have occurred yet). Safety management is therefore at the same time not only a sub task of quality management but also a wider task.

4.4.2 Operator support

Safety and quality of service imply different boundaries to operator behaviour. As is described in the alarm clock-model, operator attention will be focussed on those information sources where relevant changes bringing the system state closer to one of the critical boundaries are most likely to happen (van Westrenen 1999). This is determined by an expectation, which is based on previous observations. If variable values change only slowly or if they are far from their critical boundary values, it will take longer until this variable will receive attention again. Thus “expectations provide a way to determine what changes are informative and therefore worth practitioners’ attention” (Woods and Hollnagel 2006). This attention allocation mechanism may also explain why human operators are in general not very successful at performing monitoring tasks, such as looking at computer screens on which basically nothing happens that requires their attention (Bainbridge 1987; Kaber and Endsley 1997). They need support in detecting movement towards boundaries. In particular unexpected and high-priority changes must be reliably accompanied by some notification to draw the operator’s attention.

4.5 Further development of the safe envelope in a railway context

4.5.1 Railway operation in safe envelope terms

In this section and section 4.6, the safe envelope will be further clarified in a railway context, and from this lessons learned will be derived in section 4.7. In railways, the intended trajectories of all trains are laid down in the planned, network-wide timetable, which prescribes when all trains should depart, arrive and pass intermediate points along the route. These trajectories are called ‘train paths’. Because the timetable is constructed in such a way that all train paths are conflict-free, the process will stay within the operational envelope at all times if all trains stick to the trajectories prescribed in their timetables. However, when disturbances occur, execution of the intended timetable may become impossible, and the traffic controller and the train drivers have to deviate from the planned trajectories. The traffic controller may replace (parts of) the plan with another plan that better suits the specific situation; this updates the operational envelope and the desired system state therein to the current circumstances. For this to be effective, it must be communicated effectively and as completely as possible to the train driver. The next subsections will show how this can be

depicted in safe envelope terms. Afterwards in section 4.6, an example will be given of how this process can fail: a railway accident is described in terms of the safe envelope.

4.5.2 2D visual example in railways

Defining the system under consideration
The safe envelope will now be elucidated with an example of a train movement. The system under consideration is one train (not the passengers in it or the whole railway system with many trains) and the intended controller of this system is the driver. He must operate the train such that the timetabled service (a movement from one position in time and space to another position in time and space) can be provided and he must also keep the (kinetic) energy that the train carries within acceptable boundaries. Since the mass of the train is given, he can only do so by changing the train's speed. Therefore, time, position and speed are important variables to consider.

Defining the envelope's cross-section
The safe envelope of operations is multidimensional, but in order to visualise it on 2D paper two dimensions must be chosen. This is an inherent simplification and it should be explicitly stated that the following pictures *are* not the safe envelope, but are *a view on* the safe envelope. The train movement is drawn in Figure 4.3, which represents a distance-speed-diagram. Some other variables, such as time or the deceleration or acceleration rate, are included in the safe envelope of operations, but the two-dimensional picture can only include them in an implicit way. Other variables, such as the clearance gauge or track layout, are perpendicular to the selected dimensions and cannot be seen in this view; but they could be made visible in another. Depending on the purpose of the analysis, other cross-sections of the system state will be necessary.

The viable envelope
The viable envelope consists of the following boundaries. As soon as the train's position and the position of another train or object (the track occupation) overlap, there will be a collision and kinetic energy will be unintentionally transferred to the track occupant. This results in a

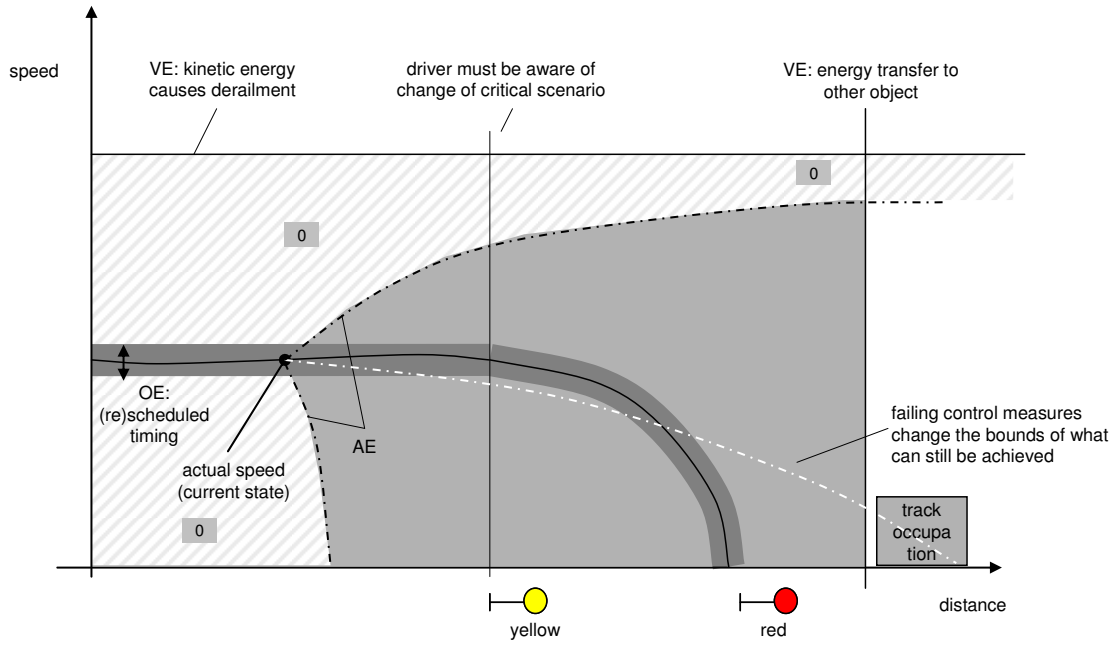


Figure 4.3: Attainable system states within the viable envelope at one point in time (cross-section of two variables).

vertical restriction (perpendicular to the distance axis) in Figure 4.3. If the train's speed becomes too high, the dynamic forces between rails and wheels might become so high that one of them gets damaged, or the train might derail. This imposes a horizontal restriction (perpendicular to the speed axis).

Note that the law of conservation of energy means that the hazard can be substituted. For instance, when the train brakes it will reduce its speed but its brakes will become hot, which can lead to other damage processes not modelled in detail here.

The attainability envelope
The attainability envelope is formed by the maximum possible acceleration and deceleration rate. The unattainable states are hatched and marked with '0'. The desired train path as prescribed by the timetable and possibly modified by the traffic control centre is given as an operational envelope. The train should be brought to a halt in front of the signal showing a red aspect, because there is a track occupation beyond it. This situation is attainable because the path towards it is completely within the attainability envelope.

In some cases, e.g. if the brakes of some cars are put out of service (white dash-dotted line), fewer system states can be attained. The attainability envelope shrinks and indicates that the process cannot remain within the viable envelope; the system state can no longer be (sufficiently) controlled by the intended controller of the system. The evolution of the system state in that case depends on outside help and random factors. Examples of the latter are late running of another train or route setting such that braking is not required. Outside help can be provided for instance when someone can move the track occupation away. If outside help is required, communication is necessary. The process might also become controllable again (e.g. restored brake functioning) and if the recovered controls are still sufficient to steer clear of the viable boundary, no damage will occur. All these scenarios can be modelled as changes to the viable envelope and the controllability envelope. Note that no accident has happened until the viable envelope has been exceeded, but once this is the case, damage is (by definition) inevitable.

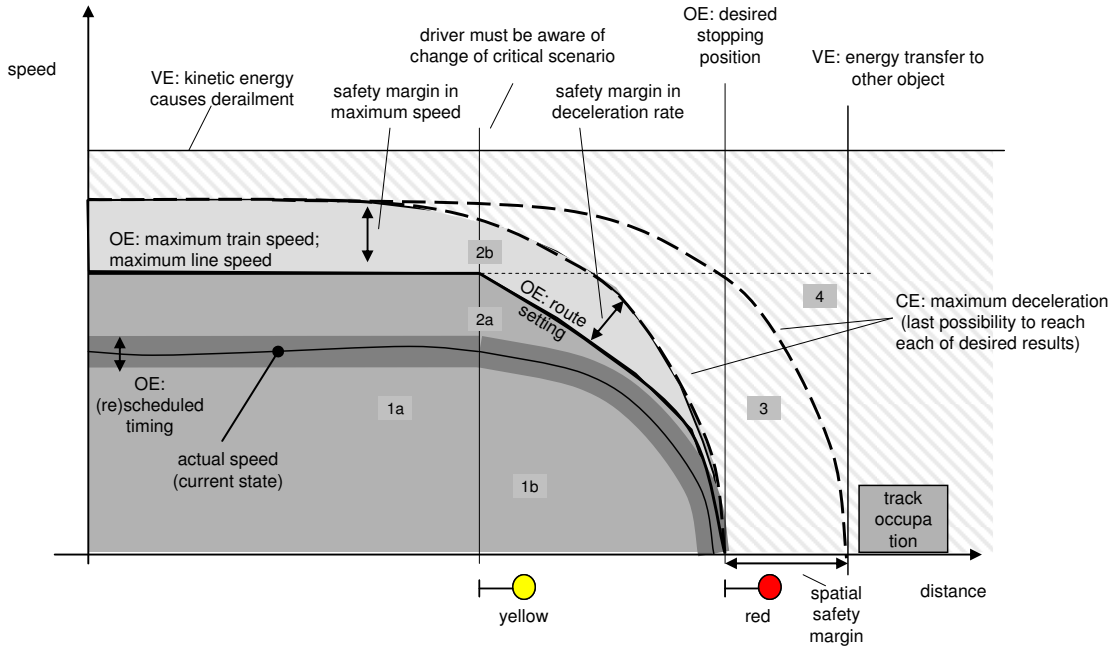


Figure 4.4: Controllable system states within the viable envelope at one point in time (cross-section of two variables).

The controllability envelope

The driver is the intended controller of the train. In Figure 4.4 we see the controllability envelope (CE), a bold dashed line that indicates at which point the last possibility exists where control can be effective to reach the desired state. We see two of these lines, one for the situation where the desired stopping location is the red signal and one for the case where the desired stopping location is the track occupation itself. The hatched areas 3 and 4 are uncontrollable and are thus areas where the system should normally not be operated. The difference between area 3 and 4 is that in the first case, the train cannot be stopped before the red signal, but still before the track occupation. In area 4, the track occupation will be hit unless it is somehow removed in time.

Operational envelopes

To help the driver to navigate through the state space, operational envelopes (OEs) are defined by several preceding functions, such as the rolling stock design, timetable design and infrastructure design. Margins are included by the various designers; they can be found between the operational envelope and the control envelope. Not all designers limit the same variables. The train's maximum speed and the maximum allowable speed on the line are the same in this example and are represented by a horizontal line; at the point where the signal shows a yellow aspect, this boundary is no longer relevant as the route setting by the traffic controller implies another operational boundary, which forces the driver to slow down and prepare to stop. We see that some operational envelopes are more important than others. The same change is also reflected in the dark grey zone which indicates the desired timing of the train, as it was planned by the timetable designer and changed by the traffic controller to fit the actual circumstances. This can only be displayed implicitly in a distance-speed diagram; a time-distance diagram (as is common in timetable design) would be more appropriate here, but then, the speed restrictions could not be modelled. This shows once more that a two-dimensional representation of the multi-dimensional envelope is an oversimplification.

Driver behaviour

A main advantage of the safe envelope model is that it can not only indicate control measures such as safety margins, but also visualise the situation and the desired or actual operator

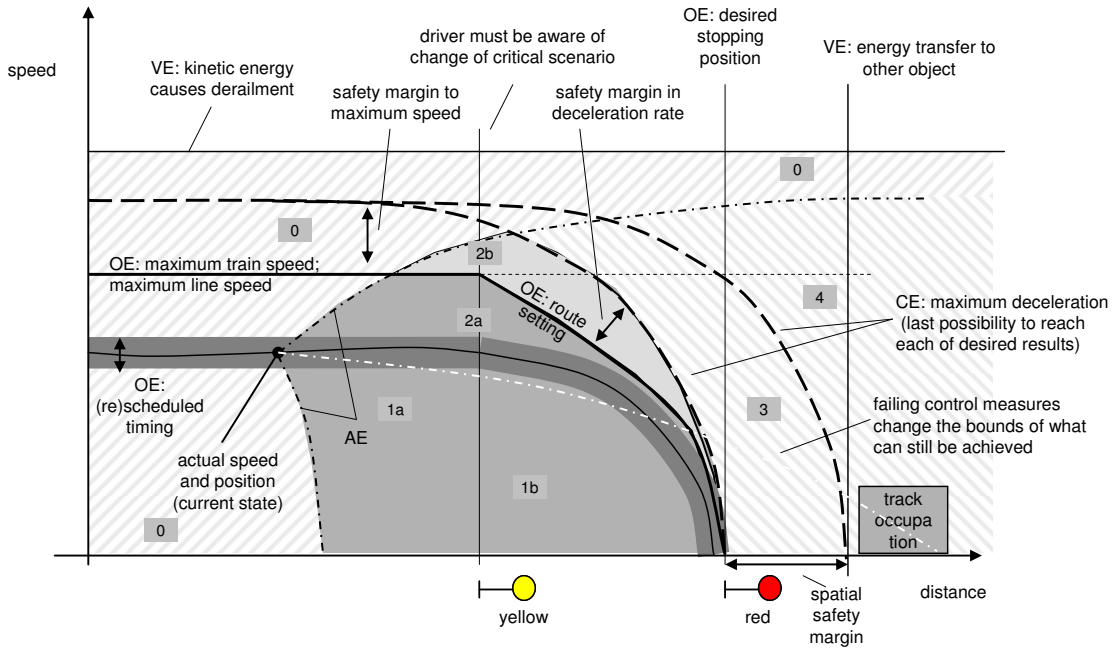


Figure 4.5: Cross-section of the safe envelope model, shown as a distance-speed diagram for a certain point in time.

behaviour. The operator will, or must, behave differently depending on which part of the diagram he is moving in. When the boundary of the control envelope is far away, it will not receive much attention and the operator's attention will be focused on other tasks, such as achieving the timetable goal (area 1a). In areas 2a and 2b, the boundaries of the operational envelope and the control envelope are approaching. Now, keeping control becomes the primary objective, so that the viable envelope will not be exceeded. The task of steering to the optimum within the envelope (for correct running of the timetable) becomes of secondary importance. In area 2a the driver stays within the operational boundary with a safety margin. In area 2b, the safety margin is used by going closer to the control envelope boundary, for example to catch up time. This will be more likely when the operational envelope is felt to be unnecessarily restrictive. This has the risk that all safety margins are used up.

Figures 4.3 and 4.4 are combined in Figure 4.5: the grey areas are both attainable and controllable. The darker the area is shaded, the more criteria from the operational envelopes are met. The hatched areas are either not attainable or not controllable, or both. The operator must be well aware of changes to the situation, especially when this implies that he should change his priorities from optimising system behaviour in order to meet the timetable to managing to stay within the control envelope. This change in priority is indicated by the vertical line separating areas 1a and 1b in Figures 4.4 and 4.5. In this example, the change is marked with a yellow signal aspect; but other indications such as cab signalling can be used for the same purpose as well.

4.5.3 Information need

As a conclusion, the safe envelope model shows that it is necessary to indicate to the controller of the system, for effective control:

- What the current system state is;
- Which goal the operator should try to achieve or at least within which boundaries he should operate;
- Whether this goal can be reached, given the available control measures;
- What the ultimate moment is to exercise control;
- Where the safety-critical boundary of operation is.

The safe envelope model can then also be used as a tool that can help designers to define which information an operator should be provided with.

4.5.4 Accident mechanisms

With its possibility to display state spaces and the envelopes, the model can reveal additional accident mechanisms compared to existing models. The fact that the system state, the envelopes and their representation in the mind of the operator can all change makes the model dynamic and allows finding more and other accident mechanisms than the sequential or epidemiological approach. Examples are:

- A change in one of the operational envelopes due to new operating procedures, changes in laws etc., so that the operational envelopes gradually allow more states than are controllable;
- Envelopes that are incorrectly perceived by operators, so that the operator thinks he can control the system whilst he in fact cannot;
- Operators who gradually tend to operate closer to envelope boundaries;
- Changes in the attainability envelope and controllability envelope: e.g. decreasing knowledge or skills of operators, wear of material, or temporary unavailability of control measures.

4.6 Illustration of an accident in the safe envelope model

4.6.1 Aims of this illustration

This section illustrates that the safe envelope model can be applied to a real accident. The chosen example is one that happened in Eindhoven in 1992 (Spoorwegongevallenraad 1993). It shows how several of the envelopes elements change during the course of the accident. Some lessons learned by applying the model will be described in section 4.7.

4.6.2 Setting of the accident

A driver of a local train was given permission to depart from the station platform, but due to a delayed intercity train he could not proceed across a junction ahead. After departure he had to stop at the next signal and wait until the delayed train cleared the crossing. He failed to do so, resulting in a collision with the intercity. This accident was chosen as it occurred during regular railway operation, without major disturbances. Yet, the accident could happen, and a series of similar accidents has happened since. The accident will be described below in terms of how the process and the envelopes move. The letters indicate subsequent phases in the accident.

4.6.3 Displaying actions by the local train driver

For the driver of the departing local train, the steps in the scenario are described below and are shown in Figure 4.6.

A. The driver saw a yellow signal aspect, which allowed him to depart from the platform, and so he did. This action is within the operational envelope that was provided by the traffic controller. However, since green signal aspects are assumed during timetable design to be available to trains departing stations, the process does not fulfil the timetable designer's specification of the operational envelope.
B. The driver did not know the operational envelope that resulted from the rescheduling by the traffic controller. The driver did not know the traffic controller's plan, and thus, he could not anticipate a proper driving style. Then, his attention was distracted and because the tracks were in a curve, he wrongly believed that a yellow signal aspect on a neighbouring track was meant for his train. Due to this wrong detection of information, he thought that he was allowed to proceed and failed to take the correct action (braking). Here, the process deviates from the intended scenario (grey band).
C. The driver, who thought that he had now been given another yellow signal aspect, therefore did not notice that his train was passing the next signal while it was showing red. The red signal, by definition, is located at the limit of the operational envelope (because passing it is not allowed) and here the process exceeds the operational envelope that resulted from the route setting process.
D. The driver, in theory, still has control over the situation (e.g. by applying emergency brakes) and therefore is within the controllability envelope. However, as no further information was presented to him from which he could have detected his error and could have noticed that he was operating outside the operational envelope, he did not take any action.
E. The driver then noticed that the points were not set correctly for his train. He then realised that he must have missed a signal at danger (i.e. that he was already operating outside the operational envelope) and applied emergency braking. As of this moment, he could not influence the train speed any further and therefore he was temporarily out of control.
F. The train came to a halt on the points, which became damaged. This meant that the viable envelope was crossed as far as the relationship between the train and the infrastructure was concerned. The local train now was also in a position that undesired transfer of energy between the two trains became possible, but there was no collision yet: the train

was still within the viable envelope as far as energy transfer with other trains is considered. The driver understood the situation and he wanted to clear the track for the intercity which was approaching with 100 km/h (which meant that his and the other train's viable envelopes were also approaching each other with that speed). He tried to drive his train a few meters back. This was not possible as the brakes were still applied and it would take 10 to 15 seconds to release them. We see here that the driver did not have control: he could not influence one of his essential variables and could not get back into the operational envelope.

K. The front of the approaching intercity train constituted the boundary for the viable envelope. It was moving inwards until it would reach the local train, unless it were stopped by the driver of the intercity train. As the latter did not occur, the accident happened.

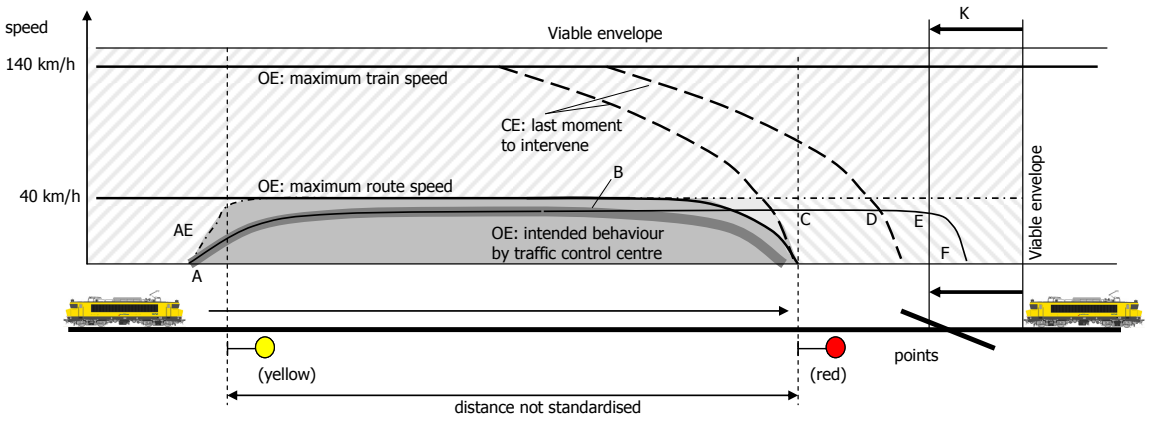


Figure 4.6: Safe envelope example for the local train.

A strong point of the safe envelope model is its ability to visualise variables, their boundaries and scenarios that show how the system state evolves. Figure 4.6 showed the actions of the local train driver in a distance-speed profile. However, this representation also has some drawbacks:

- It appears that the driver had full information, not only about the current situation but also about the future. Of course, he did not have all of the information that is shown in the picture, such as the precise location of the controllability envelopes, or the location of point B, where he unintentionally executed another course of events than what was required.
- It shows only one scenario. This is fine for retrospective analysis, but the picture would change dynamically if the traffic controller or the driver had taken other decisions.
- An inherent weakness to the use of two-dimensional paper is that it cannot show all relevant variables at once, and so a number of variables may be forgotten in whichever projection is chosen.

4.6.4 Displaying actions by the intercity driver

We will show the actions of the intercity train next, but in a different, more abstract representation. Each picture can be considered as a slice of time, representing all variables in an abstract way. The attainability envelope, which is defined as the system's ability to reach a certain state at some point in the future, cannot be drawn in every time slice, because it is only meaningful if such a future state can somehow be defined. The attainability envelope is only drawn where it is useful to do so. These figures interpolate between representations F and K in the local train driver's view described in section 4.6.3.

G The intercity driver was trying to catch up a small delay. This was within the operational boundary; but because of the delay, the current system state deviated a little from the optimum.

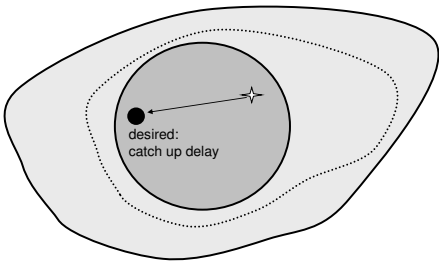


Figure 4.7: Intercity train's envelope (G).

H The local train, that had passed the signal at danger (see step C), came into the route that was assigned to the intercity train by the traffic controller. For the envelope of the intercity train, this was an inwards move of the viable envelope, as a position within its operational envelope became suddenly an unsafe place. This was also indicated by the signalling system, which detected that the route that was assigned to the intercity was no longer free and the signal that gave access to the section in which the local train now stood, automatically showed red.

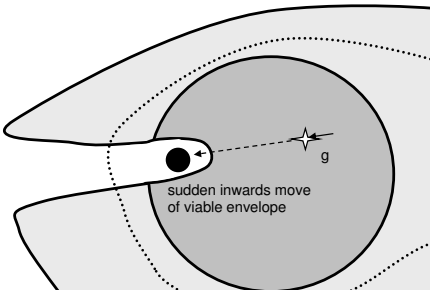


Figure 4.8: Intercity train's envelope (H).

I The former goal was now outside the viable envelope, which was undesired and therefore could not be the optimal system state. The goal suddenly bounced back to within the viable envelope. This meant that, all of a sudden, no (or fewer) effective control measures were in place to keep the system within the new boundary: the new goal was outside the attainability envelope and the controllability envelope became smaller, thus indicating that the situation, which had been under full control of the driver, had now changed such that no effective control measures were left.

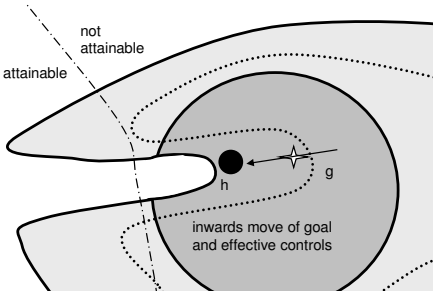


Figure 4.9: Intercity train's envelope (I).

J The intercity driver noticed that a train was standing in his route. He knew that he would not stop in time to avoid exceeding the viable envelope; there was no attainable system state within the viable envelope. Note that he was, at that very moment, still within the viable envelope of the intercity train, as no damage had occurred yet. However, no control measure was powerful enough to keep the train inside it in the very near future. All that the driver could do was to use the emergency brakes, which served two purposes: it reduced the impact energy and therefore the damage and it created time for the local train to regain control. In the viable envelope of the local train the moment of the impact was a little bit postponed. Thus, whilst the intercity driver had no more control, he tried to give time to other actors (especially the train driver of the local train) to solve the problem.

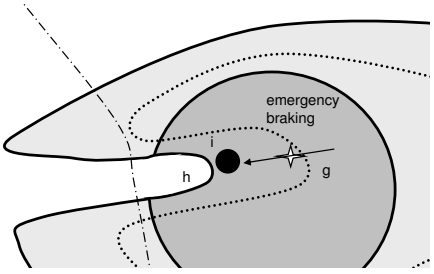


Figure 4.10: Intercity train's envelope (J).

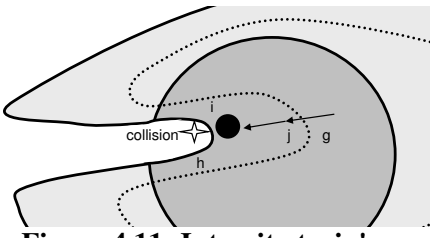


Figure 4.11: Intercity train's envelope (K).

K For both trains, there was a final event: as neither of the drivers had effective control measures left, the two trains collided, thus passing their respective safe envelope boundaries, and damage occurred. 48 people were injured and both trains were heavily damaged.

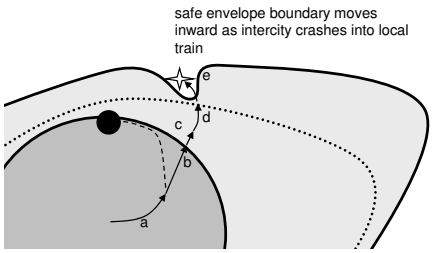


Figure 4.12: Local train's envelope (K).

This representation is somewhat unspecific about exactly which variables are concerned. This invites the user to think about which variables are important, and this mitigates the risk that only the displayed variables are taken into consideration. Time is perpendicular to this representation. This has both advantages and disadvantages.

- The disadvantage is that the attainability envelope cannot be drawn easily, because this is defined for a certain point in the future and can therefore only be shown if it is indicated for which point in the future it is meant or if certain states are unattainable for any relevant point in time, such as in phase I;
- The advantage is that this representation does not show any information that the operator also could not have.

If the goal is to display attainability, the display mode of Figure 4.6 is more useful. If the goal is to show on which basis a driver can take decisions, a display mode similar to that in Figures 4.7 – 4.12 can be more appropriate. This is not so much a limitation of the safe envelope model, as it supports both views, but rather a limitation of displaying dynamic information in static two-dimensional pictures where in fact more dimensions are relevant. One recommendation of this research is to further look for support to enhance displaying the right parts of the multi-dimensional envelope.

4.7 Lessons learned from applying the safe envelope model

4.7.1 Possibility to model intended system behaviour

This section will reflect on the safe envelope model and its ability to inform the discussion on a number of issues in railway safety. The Eindhoven accident will be used as an illustration, where appropriate. In the example, the safe envelope has been used to depict the accident. However, the model also allows the display of the intended system behaviour. This same sequence of events was carried out frequently in Eindhoven and generally did not lead to an accident. The local train then stopped at the red signal. The intercity would pass first and the operational boundary for the local train would then be enlarged when a route for this train was set.

The formal accident investigation report largely focussed on improved automatic train protection, so as to stop the train in the case of driver error. The safe envelope model indicates that this could be an effective solution, but that also provision of additional information to the driver on his desired behaviour (between points B and D in Figure 4.6) may be helpful in order to achieve the desired system state and to stay away from the viable boundary. Examples of such improvements could be measures to help the driver to detect the correct signals for his route, or information from the traffic controller about his intention about the train order on the crossing, or the time that the local train is expected to be able to pass. Such information measures would not only make the situation safe(r) but would also enhance process control under normal circumstances. However, the benefits of such extra information

are difficult to envisage when safety, and thus the avoidance of undesired situations, is the main viewpoint, instead of the question of how to achieve what is wanted in a safe way.

4.7.2 Effectiveness of control measures

The measures available for control are extremely important for staying within the envelope boundaries. One reason that the Eindhoven accident became an accident was because the intercity did not have enough time (given its speed, its braking power and the remaining distance) to come to a stop before crashing into the local train. If the intercity driver had had better braking equipment or if he could have been aware of the local train standing in his route some 200 meters earlier, the collision would not have happened because the controls would have been sufficient. Also, if the local train driver could have released the brakes and backed up his train[6] more quickly, the collision might have been avoided.

The question now is: when does control become ineffective? The simple answer is that, if things change more rapidly than the controller can keep up with, control will be ineffective. A sharper definition for the controllability envelope will then be: 'if the operator does not take action at this point, no measures are left to ensure the desired result, i.e. to remain within the viable envelope'. Formulated in a more mathematical way, one could say that all possible (combinations of) control measures R need a certain time t for the desired effect to be reached (see Table 4.1). For effective control there should be at least one measure for which this time t is smaller than the available time T, within which the result should be achieved, because otherwise the controllability envelope will be crossed. If this last measure is not chosen ultimately at the moment where t equals T, no effective control measures will be left, as also the last measure then has a $t > T$. In that case, all measures require more time than is available. Figure 4.9 shows that the controllability boundary suddenly shrank, because the available time for action was reduced. No measures were left that could be completed within this time and the measures that could be taken could serve only to reduce the impact of the accident.

The available control measures and their required time to take effect define the attainability envelope. The controllability boundary is defined by the last (combination of) measures that can just deliver the required system state within the required time T. In Table 4.1, R_2 is the only effective measure and it should be applied immediately. If there is only one possible way left to reach the desired system state in time, this also means that one single error (or even doubt about which measure to take) will sooner or later result in an accident.

Table 4.1: Different control measures R have different times t to take effect.

Control measure (attainability)	Required time		Available time (controllability)
R1	t1	>	T
R2	t2	=	T
R3	t3	>	T

4.7.3 Operator awareness

There are several situations in which control will be ineffective and certain states will be unattainable:

[6] Note that the railway's control philosophy in principle does not allow for backing up trains. This can lead to derailments or the braking distance of the following train may be reduced whilst its driver is not warned in time; yet to avoid a certain accident, taking such a risk may still be beneficial.

- The operator did not realise that he should use a control measure (wrong estimate of T);
- The operator did consider a measure, but rejected it incorrectly in the belief that it would not be helpful (wrong estimate of t);
- The operator incorrectly chose a measure that will not reach the desired effect (wrong R);
- The operator incorrectly chose a measure that will not reach the desired effect in time ($t > T$);
- The operator did not consider a measure as an appropriate one (R forgotten);
- All available measures take too much time to reach the desired effect ($t > T$);
- The required measures are (temporarily) unavailable (no R or $t > T$).

In other words, effective control requires information, a correct representation of the system state in the mind of the controller, and available control measures. If one of these is lacking, no control can be exercised and the system state cannot be influenced properly. By using the safe envelope model an a-priori analysis can be made of whether such situations may occur and how design modifications can prevent such situations or make them controllable. Whether control was effective can only be determined afterwards, as it requires the result of the measure to be known. For example, the intercity train in the Eindhoven example applied the emergency brakes. Because the exact braking distance is not known upfront, this could be either a control measure or a mitigating measure, depending on whether it could serve to prevent the accident or just to decrease the severity of the consequences. Which of the two it was, will only become clear after the event. It is therefore impossible to draw the controllability boundary and the attainability boundary with absolute certainty in advance. They will always be an estimate of what will be effective and what will not. Experience, education and/or research may help people to have a mental model of the situation and make better predictions about the system's controllability[7]. Having such a (mental or computerised) model is essential to control a system, as Conant and Ashby (1970) pointed out: without a model, the controller would not be able to decide on an effective control measure. He has to map the system state into his own model, which can also be a source of control errors. Figure 4.13 shows several of such mapping errors, leading to a representation in which the actual state appears to be within the operational envelope, where it in fact is not. The controller also has to make an estimation of possible measures R, their required time t, and the available time T: the driver of the local train in Eindhoven thought that he was in full control. He did not take the required action, because his mental picture did not match the real world. When he noticed information that forced him to adjust his mental picture, no time was left for an effective control measure.

4.7.4 Regaining control

Control may be regained when more time is given, because control measures may then become available that have their effect within the available time T. Another possibility is to call in the help of another actor: he may have control measures for which t < T or may be able to change the situation and thus shift the viable envelope, thus increasing the available time T. The driver of the local train, who was out of his own control envelope, could have made an emergency call so as to try to initiate a control measure in the approaching intercity's envelope. If the intercity train in the example had been further away, its driver would then have been able to stop in time.

[7] We also have to face the dilemma here that better controllability of the system may tempt the operator to steer the system closer towards the operational envelope, making the system more vulnerable to the remaining uncertainties. This operationalises Wilde's risk homeostasis theory and Rasmussen's migration towards the boundary of acceptable performance.

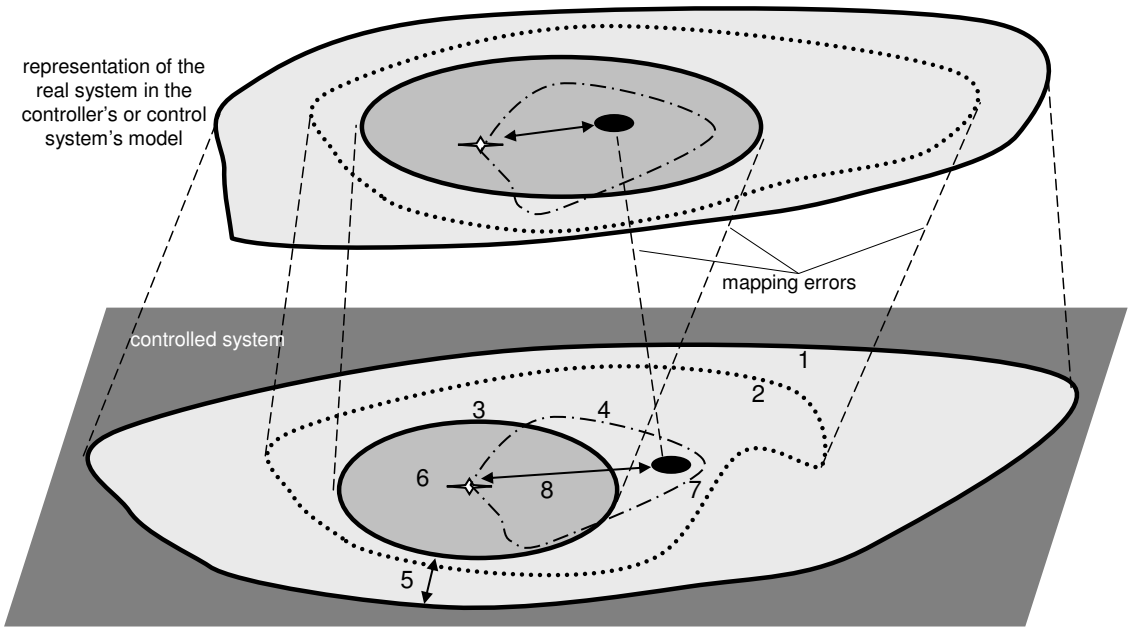


Figure 4.13: Mapping of real variables to a representation in the model of the control system.

4.7.5 Multiple controllers in one system

In railways, there are multiple controllers, such as the traffic controller and the drivers in his area. None of these controllers has a complete overview of the whole system, so that these controllers must co-operate and exchange information.

When an actor has lost control, accidents are not inevitable if other actors can still change the situation. An accident only becomes inevitable when none of the actors have effective control measures left. This emphasises that the envelopes must be modelled for each actor separately, but they do interact as we saw in Figures 4.7 – 4.12, where the presence of one train modified the viable envelope of the other one. Also during normal operations, in which no actor has lost control, separate system functions are often controlled by different controllers. These have to cooperate in such a way that the states of each of the subsystems will always remain within the respective operational envelopes. Their cooperation requires that the controllers somehow possess accurate information about the controllability envelopes of each of the other controllers they interact with, so that no controller will manipulate the system's state to bring it outside the controlled area of one or more of the partners.

Not only are there many drivers and traffic controllers, but there are also designers and managers. They may have completely different interpretations of each other's work. For an elaborate classification refer to Busby and Hibberd (2002). Since each of the actors has a specific task in the ensemble, they will all have different models and thus different representations of the same envelope. Operational rules are supposed to ensure this matching of perceived envelopes. But also the developer of the rules will have his own view of the system, which needs to match reality, but may not if he is not informed by experience and feedback. Absence of required communication creates problems. For example, traffic controllers currently assume (and are also taught) that drivers are fully informed by their signal aspects, which however is not the case, as a systems-theoretic analysis of signalling (Chapter 7) and also interviews with drivers revealed. In some cases drivers need additional information from the traffic controller to provide a clue about where the envelopes and the optimal goal state within them are, but currently they do not get this as a matter of course.

4.8 Discussion

4.8.1 Contribution

This chapter served to further develop the initial safe envelope model as proposed by Hale et.al. (2007). The following contributions have been made:

- The terms and definitions used in the safe envelope model have been listed clearly;
- The model was extended with the attainability envelope and the production-related issues, such as the goal state and the deviation from that;
- Thus safety and production could be combined in one model;
- Detailed examples of how the safe envelope can be used to describe operator behaviour and accident mechanisms were given;
- The model enables designers to derive the main information need of the operator controlling the system. (Perhaps dedicated human factors models could further expand this list);
- A connection was made between the safe envelope model and the awareness of operators controlling the system (this will be further elaborated also in the next chapter);
- The safe envelope makes clear within which boundaries the process should be controlled. Combination of the safe envelope with a control model is required to take advantage of the full benefits.

4.8.2 Fulfilment of requirements

In section 3.2.7 a number of requirements for a new risk control model were formulated. This section will discuss to what extent the safe envelope of operations meets these criteria:

1. *The model should describe the operational level with all its relevant variables, in which operators are controlling the events in the system. Given the design and organisation of the system, these are the people who are the last to balance the risks and the benefits of the activity. They can decide how the activity will progress.*
   Operations can be seen as reaching a certain goal in time; in other words as making certain variables conform to a desired value. If we want to better support real time operations, we need to understand how this process works. The safe envelope model focuses on processes that result in changed values of the system's characteristic variables. It is not relevant whether a variable was changed by a human operator or by an automated controller or by external events; all can be modelled. Also interaction between operators can be modelled: these are reflected as changes in one of the envelopes. Such changes may remain unperceived without information flows to alert the controller about the change. Therefore, communication is required, a fact underlined by the model.
2. *The model must be able to represent quality of service (the goal that the operator should meet);* and
3. *The model must be able to represent the energy contained in the system and in its environment (in order to model accidents)*
   The model can show, within its multi-dimensional variable space, how scenarios are developing; be it a desired scenario leading towards the goal or an undesired, deviating scenario leading to an accident or at least not leading to the desired goal. Quality of service imposes constraints to certain states that are undesired; the need to regulate the energy in the system imposes constraints to, for instance, the physical controllability and attainability (e.g. deceleration rate by changing kinetic energy into another form of energy). Both require the system to be controlled. Thus, safety is not modelled in isolation, but in the context of the production goals that have to be pursued. This seems to be an essential feature as the context in which decisions are made plays an important role in how the decision is being made.

4. *The model must be able to reflect earlier choices, such as in the design and management level. These define and influence the conditions in which operators have to do their work, such as operating procedures, technical aids and the means that they have at their disposal to control system processes;*
   Design and management decisions can be depicted in the safe envelope model; the outcome of their decisions can be shown as operational, viable and controllability envelopes and as the position of the goal state. It is also a management task to make sure that controls are provided and kept in a working order, thus maintaining the controllability and viable envelopes intact, where they are expected to be at any one time.
5. *The model must be able to respect relevant influences from outside the defined system boundary;*
   Such influences can be incorporated as (external) processes that may affect control measures or variable values, or may add or change boundaries in the envelope.

We conclude that the safe envelope model does now meet the requirements: it can therefore contribute to our understanding of which mechanisms cause risks to arise and how these can be controlled. This is an important first step, that tells us which things are important and why.

A subsequent step, but one beyond the scope of this thesis, could be to model these things in a more formal mathematical way where possible. However, this can only be done after a systematic analysis has shown which relationships must be mathematicised, i.e. which mechanisms underlie accidents. The safe envelope model can help to distinguish these. At this point, we should warn that not all relationships can easily be quantified and that quantification is not an end in itself, but it may help to describe things more explicitly wherever this is useful. Another next step could be to think of ways to improve the visualisation of the multidimensional envelope, for instance by developing software that can render cross-sections of various variables or 3D-pictures (although this may still be fewer dimensions than required), that can select the most urgent 2D pictures or that can show the dynamics of the way in which the envelopes change.

4.9 Conclusion

The safe envelope model allows us to see which control measures and information are necessary in order to control the system properly, which means that it can take both safety and quality of service into account. Actors controlling a system should be able to see:
- What is the current system state?
- What is the goal and which tolerances are allowed (operational envelope)?
- Can the goal be reached in time (attainability envelope)?
- Where is the boundary of safe operation (viable envelope)?
- What is the ultimate moment to exercise control (controllability envelope)?

In distributed systems such as railways, the answer to these questions generally requires communication between different actors, which is therefore of vital importance to keep the system controllable and safe. The safe envelope model also makes clear that drivers and traffic controllers not only need aids to indicate what they are allowed to do, but also aids to navigate in the dynamic environment in which they are operating. Such aids could for example provide additional information about expected conflicts in rail traffic, and thus guide a driver's or a traffic controller's attention towards those parts of the process that need their attention to solve the conflict. With increasing network utilisation, there will be more dynamic interactions and thus more need for such decision aids. Better control may then improve

safety and quality problems. The next chapter will provide more information about control, and thus serves to further enhance the safe envelope of operations.

5 Process control within the safe envelope

5.1 Introduction

The safe envelope of operations, which was introduced in the previous chapter, defined boundaries within which the process should remain, and showed that the boundaries and the process therein are dynamic: they can change over time. It was stated that it requires control to steer the process such that it remains within the boundaries. This chapter deals with the question of what is required to do that. It will first give a short overview of existing process control theory, based on cybernetics, in section 5.2. Following that, we will adopt and further develop the DDA (Detect – Diagnose – Act) method from safety science and associated with the safe envelope model (Hale et al. 2007). In section 5.4, a connection is made to models from human factors, management science and safety science, where DDA loops can also be found. Section 5.5 provides a discussion, in which the Eindhoven accident from the previous chapter is re-described in terms of DDA loops.

5.2 Cybernetics

Process control in this thesis takes cybernetics as a starting point. Cybernetics is the general science of controlling systems. It focuses on functions and information flows rather than on what the system consists of, and it can therefore be applied to a broad diversity of systems, such as business, technical and biological systems (Heylighen and Joslyn 2001). Cybernetics is primarily aimed at the general principles of control. The cybernetic principles are compatible with those of process control theory, which has the objective 'to maximize production while maintaining a desired level of product quality and safety and making the process more economical' (Hahn and Edgar 2002). Process control theory focuses mostly on technical systems and can therefore use mathematical equations to describe system behaviour. This is in itself advantageous, but it bears the risk that those elements that cannot easily be quantified are not included in the model.

Any system's behaviour can be described by making use of five basic terms, which are depicted in Figure 5.1 and are described below (Ashby 1956; Heylighen and Joslyn 2001):

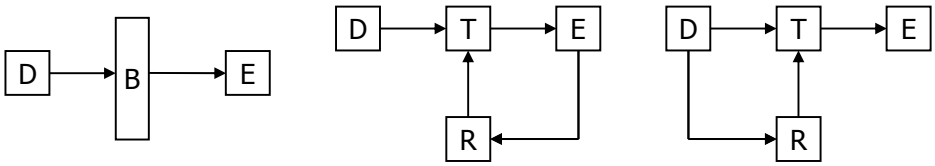


Figure 5.1: Three basic types of regulation: buffering, feedforward and feedback.

E In order to survive or to be run satisfactorily, the system needs to reach a certain goal: the essential variables E should be within the range of their desired values (η). Ashby, one of the founders of cybernetics, stated: 'These two must be given before all else. Before any regulation can be undertaken or even discussed, we must know what is important and what is wanted' (Ashby 1956). There are two forces that influence the value of E:

D, R Disturbances affect the essential variables, possibly forcing them out of their viable or operational range. The Regulator (also called control system) tries to maintain the essential variables within their viable range.

T T is the transformation process in which D and R affect the state of E; it is the operational process in which interactions materialise.

B A buffer is a fixed measure (i.e. one that is not chosen during real time control) that can absorb disturbances to a certain extent so that they do not affect E. Buffers are useful to handle disturbances which have been chosen not to be controlled or are deemed uncontrollable.

Control theory in engineering differentiates between control systems that are designed to reach a constant value and those where the goal is to follow a certain path (e.g a train path). The first are called regulator, the latter tracking or servo systems (Hahn and Edgar 2002; Franklin et al. 2006). This difference is not made here, where both types are regarded as a regulator in cybernetic terms. In the examples given in this thesis, both types can easily be converted into each other, for instance by subtracting the planned train path from the realised one (which results in train delay) and trying to keep this value close to zero. The servo mechanism now has become a static regulator.

5.2.1 Three control types to counteract disturbances

In order to make the essential variables E correspond to a value in their desired range η, there are three generic ways to exercise control (Heylighen and Joslyn 2001), which are shown in Figure 5.1:

1. *Open loop control*
   This is a passive control mode, in which disturbances are estimated during design (so before they are actually there), and control measures are taken to prevent the expected disturbances from influencing the essential variables. Buffers can be used to absorb the disturbance. The size of these buffers is often determined during the design of a system (e.g. buffer times during timetable design). The advantage is that there is no need to think of a countermeasure for every disturbance, because the countermeasure is already predefined. It therefore works also in such cases where the disturbance or the essential variable values cannot be easily measured. The drawback is, for the same reasons, that one cannot be sure that open loop control will be effective. An example is a washing machine, which does not measure whether the laundry is clean before it stops.
2. *Feedback control*
   This is a reactive control mode. The regulator R measures how much the essential variables deviate from their desired values and it creates a corrective action once a deviation passes a certain threshold. This has as an advantage that the deviation is known precisely, but small deviations will always remain, since control can only be exercised

once a deviation is present. Also the rate with which a feedback controller can change the system, should be higher than that of the disturbances, as the deviation will otherwise grow despite the regulator's activity. An example is a thermostat.

3. *Feedforward control*
   This is an anticipatory active control mode. The regulator R measures the disturbance D, estimates its effect on E and chooses a measure that influences T such that no net effect on E will be noticeable. An advantage is that there is time to act before the disturbance influences the essential variables, so that control can be perfect in theory. However, this requires perfect knowledge of all disturbances and a perfect model that predicts how these will influence the essential variables. Also, there should still be enough time to act. Since such perfect knowledge does not exist, small errors will always remain. An example is to listen to the weather forecast and decide which clothes you need to wear.

These modes each have their specific advantages and disadvantages. They can be combined in order to get the best possible result.

5.2.2 Prerequisites for effective control: Ashby's Law of Requisite Variety

In order to effectively regulate a system, four conditions have to be fulfilled (Leveson 2002). The regulator needs to have:

- a *goal* (the viable range of E towards which the system should be steered);
- *control measures* (so that the system can be influenced in T; also known as controllability);
- a predictive *model* of the system (so that R is able to assess the effectiveness of possible control measures);
- *information* on the actual system state or the disturbances (to see where the system state E is, or where it will go, in relation to the goal state; also known as observability)

It was shown by Ashby that the regulator R, in order to reach the goal, needs as least as many countermeasures as the disturbances D have possibilities to restrain the regulator from reaching it. This is *Ashby's Law of Requisite Variety*: 'only variety in R can force down the variety due to D. Variety can destroy variety' (Ashby 1956). It is an old and generically applicable law, but one too often broken. Before applying this law to the railway's control system, we will elucidate it with an example.

5.2.3 Example: room lighting

Imagine an ordinary office room in which we want to keep the lighting at a constant level. This is the goal (E). The room has one window to the outside world, through which disturbances (D) come: e.g. changes in the position of the sun, clouds or thunderstorms will change the light intensity in the room. A regulator (R) must be in place to counteract these disturbances. It should have at least one countermeasure for every disturbance. At sunset, the regulator should be able to increase the amount of light, for instance with a light bulb. However, this light bulb has only two possible states and therefore a low variety: on or off. The light level will therefore not be at a constant level. When the regulator has a bulb with a dimmer at its disposal, it can create much more variety to counteract the gradual darkening. In the morning, the sun will shine from a different angle and flood the office with light. The regulator now needs curtains to decrease the amount of sunlight that comes into the room. However, curtains have a limited variety to counteract the sunlight. Venetian blinds are able to create the variety needed to regulate the light from the sun. In this example, the regulator will be able to reach the goal under most conditions. In some cases, the regulator will not be able to compensate all possible disturbances, such as flashes of lightning during a thunderstorm, or failure of the electrical supply.

5.2.4 Effect of time on effective control

Although it is not stressed in Ashby's original theory, the effect of time is also important, since the variety produced per unit of time must be compensated for at at least the same rate by the controller (Beer 1985). In the example, the lightning flashes too fast for the control system to react. Basically it is advantageous if a regulator can predict accurately how a situation will change as such information provides more time to react and gives the theoretical possibility to exercise optimal control (Heylighen and Joslyn 2001; Franklin et al. 2006). Franklin et. al. (2006) state: 'in any case, the control designer should study the process to see if any sensor could anticipate either a track to be followed or a disturbance to be rejected. If such a possibility is feasible, the control designer should use it to feed forward an early warning to the control system'. Appendix A provides two examples of driver assistance systems. In a Swiss example (Co-Production, Appendix A8), the train driver is provided with information over the path to be followed; in a Dutch example (RouteLint, Appendix A7), he receives information about disturbances that can be expected.

5.2.5 Representing variety in the control loop

Ashby's Law implies that the communication channel towards R (from E or D, depending on feedback or feedforward control) must be able to transmit all possible scenarios in a unique way in order for the regulator to detect the current situation. If it cannot, the regulator will have an incomplete representation of the situation and is bound to come up with an inappropriate control action. Similarly, if the communication channel between R and T has insufficient variety, the control measure cannot be transmitted uniquely and control is also imperfect (Casti 1985). With regard to transmission of control information, Von Foerster (1985) indicated that: *'Information is a relational concept that assumes meaning only when related to the cognitive structure of the observer'*. This implies that *"the user and the informational 'source' together determine the informational value and the degree of behavioural control that results"* (Corning 2001). In conclusion, it follows that information is not objective, but must always be interpreted with the properties of the receiver in mind.

In order to exercise proper control, the regulator must have a correct representation of the real system: it needs a model, regardless of whether this is a human controller or an automated one (Conant and Ashby 1970). Several issues can cause the model not to reflect the real system. Examples are measurement errors, time delay in the measured information, the fact that only certain parts of the real world are measured, or the model itself may be inaccurate or incomplete. Also, there may be a time delay until a chosen control measure takes its effect, so that the system state can change before the measure is actually applied. In all these cases, control will not be optimal and can in some cases even work counterproductively. The human controller's model may also fail to reflect reality if it relies too heavily on expectations built up by experience and does not check sufficiently that they are correct in this instance.

The controller's model and its amount of detail depend on the purpose of the model, since it will only incorporate those elements that are considered relevant (Heylighen and Joslyn 2001). A too simple model will lead to misunderstanding of essential relations, but a too complex model will not lead to the intended level of abstraction. In both cases, this will hamper a proper diagnosis of the problem under consideration. This is clearly reflected in Einstein's quote: 'Make everything as simple as possible, but not simpler'.

In the section above, we have implicitly seen three control subprocesses: detection of information by the regulator, interpretation by the regulator, and taking an action. These will be elaborated on in the next section.

5.3 Detect – Diagnose – Act: three general control subprocesses

5.3.1 From control prerequisites to three control subprocesses

In order to understand the control mechanism, we can split any control process into three generic sub processes. Given the analyses above, a control measure can be considered effective only if the following sub processes are correctly performed:

- (D*etect*) information must be present, looked for, and perceived by someone;
- (D*iagnose*) the information must be processed correctly into a meaning, the future system state must be projected and an appropriate action must be chosen;
- (A*ct*) the action must lead to a result that corresponds with the purpose (goal) of the system in time, so that the process will move towards the desired state.

This same sequence has already been empirically derived in safety science (Hale et al. 2007). In fact, these are the same requirements as mentioned by Ashby, but with a stronger emphasis on the information, the user of the information (the regulator), and the action to be taken. We can map these sub processes on the control loops that we have seen before in Figure 5.1. The result is shown in Figure 5.2.

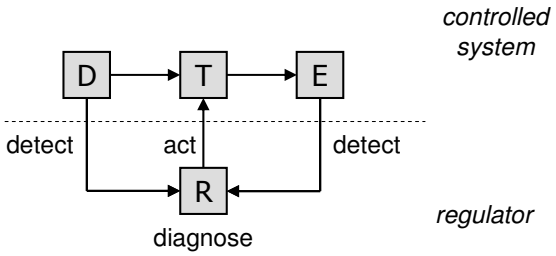


Figure 5.2: Detect – Diagnose - Act sequence mapped on cybernetic control loops.

Some remarks on similarities and differences between these approaches are given below:

- There are no explicit objectives included in the DDA cycle, which is the case because safety science is, or was, primarily concerned with the question how accidents could be prevented. In that case, the question how to reach the system's goal appears to be of less importance. However, as Chapter 3 showed, people do make trade-offs between pursuing their goals efficiently and the safety risks involved in doing that.
- There is a difference between the requirements from cybernetics and the DDA sequence in their formulation: in cybernetics, the requirements are formulated as items that one must have, such as a model or a control measure; the DDA cycle is formulated in terms of functions that have to be fulfilled. The latter has our preference, because it can well be the case that information is provided but not detected, that a model is available but no diagnosis is made, and that control measures are in place, but not used.

Therefore, in this thesis we will use the DDA sequence, but we will complement it by adding a goal. Since the DDA sequence is meant to influence the system's state, we can also combine it with the safe envelope of operations. The several operational envelopes will then act as boundaries within which the regulator should try to achieve the best possible system state. Each of the three sub processes will be elaborated on in further detail in the following sections.

5.3.2 Detect

In order for the regulator to assess the situation he wants to control, he needs to know the current system state relative to the several envelopes' boundaries, the goal that he wants to pursue within these boundaries, and whether the system goal or an approach to one of the boundaries needs attention must urgently. This was described in section 4.4 ('Balancing

attention between safety and quality‘). This means that the regulator must be able to obtain information about the envelopes, the essential variable values and, depending on feedback or feedforward control, also on possible disturbances. If there is a human operator, his expectation about the scenario will influence his perception of new information. Basically, human operators tend to interpret information as a confirmation of their expectation unless there is a strong indication that their expectation is wrong (Endsley 1999). On the other hand, technology will only take that information into account for which it was programmed, even if a failure could be detected from other information, so that the functioning of technology is ultimately also dependent on the knowledge and expectations of the system designer.

5.3.3 Diagnose

The diagnosis serves to comprehend the situation and to choose a control measure if this is required. Diagnose and detect may be an iterative process, as an initial diagnosis may generate the need to detect more information to confirm or test that. After the diagnosis of what the current situation is, an expectation is formed about the system’s future behaviour. This requires some kind of a model. Depending on the decision whether the expected situation is acceptable or not, comes the question which actions are available to steer the process to a more desirable state: either towards the goal state, or away from one of the envelope’s boundaries which the process is approaching. In order to choose the right control measure, some model is again required that will enable the regulator to determine whether the control measure will have its desired effect.

A new DDA sequence has to be started to see whether the process will evolve as expected. As opposed to machines, that only ‘look’ at variables that were specified as important by their designers, human operators will make an estimation of when and where relevant new information can be expected based on the current diagnosis (Endsley and Kiris 1995; Endsley 1999; van Westrenen 1999; Houtenbos 2008). The diagnosis thus not only serves to choose an action, but also to determine when and where information will be collected in the next DDA sequence.

5.3.4 Act

The diagnosis is followed by taking the required action correctly and in time. This can be a change to the system state, but it can also be a changed envelope boundary for one of the subsequent functions that have to be fulfilled. Of course, the decision that no measure needs to be taken can also be the final stage of a DDA sequence.

5.3.5 Control can replace buffers and barriers

According to Ashby, buffers are only one of three possible control modes as Figure 5.1 showed. This means that where currently buffer times are used to absorb dynamic effects of railway operation, one could also exercise real time control (feedforward and/or feedback) in order to keep dynamic effects as small as possible. In safety science, the concept of barriers is well known in order to prevent a system from running into an unsafe state (Haddon 1973; Johnson 1973; Hollnagel 2001). A simple example is a fence to prevent items falling off an edge. Note that the Hazard – Barrier – Target model (see section 3.3.2) has exactly the same appearance as buffering (open loop control, see section 5.2.1): the disturbance D should not reach the essential variables E and hence some barrier is placed in between. One of the disadvantages of the barrier concept is that they are easily perceived as a physical object, but they need not be. For instance, a warning sign may also warn people not to fall off the edge, but it needs active detection, interpretation and action in order to be effective. This is the detect-diagnose-act sequence we have seen before. If the item that is about to fall off the edge cannot see the sign or it cannot understand it, or it cannot take action, it will still fall. In fact,

every intervention into a system preventing it from running into an undesired state can be considered to have a barrier function. We can conclude that it must be possible to use also the other two types of regulation to provide a barrier function.

We can now also state that within his own controllability envelope, an actor is by definition able to perform a DDA sequence autonomously; outside the controllability boundary, detection and diagnosis may still be possible, but the action has to be taken by someone else and thus the only action outside the controllability boundary is to communicate the need for help, as was shown in the example in section 4.6.

5.4 Same subprocesses in other fields

5.4.1 Compatible methods in several fields

This section will show models from the fields of management, human factors and process control technology, which have a core similar to the DDA sequence. This makes it plausible that the DDA model can be used to answer the question how quality and safety in a socio-technical system like railways can be improved by enhanced process control, without drawing conclusions that are incompatible or incomparable to these fields.

5.4.2 DDA in management: the Deming circle

In section 2.3.2, it was explained that every actor in the cascade diagram performs a Deming circle (plan – do – check – adjust) (Deming 1982). This is a model well known in management. It is also known as continuous improvement and emphasises the need to check the results of an activity against the goal. Similarly as in DDA, this goal is not explicitly mentioned, possibly because it is not an activity. If there is a discrepancy between the measured result and the desired result, adjustments should be made (often referred to as Act, which is, however, confusing with the wordings of the DDA acronym and is therefore replaced here with the word 'adjust'). This process can also be mapped on cybernetics and the DDA sequence; see Figure 5.3.

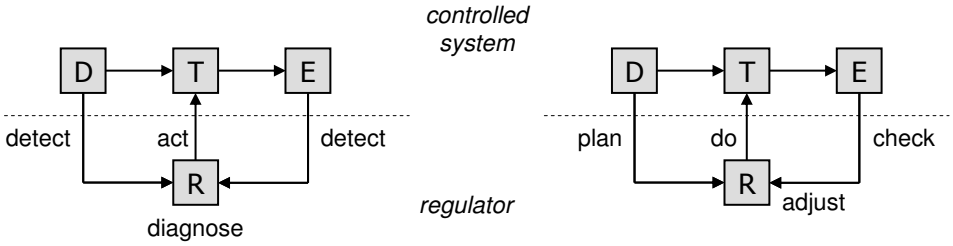


Figure 5.3: The Deming circle and the DDA sequence related to cybernetics.

Thus, DDA cycles do not only function for real time operators but also on other levels in the cascade model. For example, the task of the railway inspectorate is to retrieve information, interpret this information, and intervene if necessary (see section 2.2.8), which clearly is also a DDA loop.

5.4.3 DDA in human factors: situation awareness

Situation Awareness (Endsley 1995) is a concept used to understand how human operators control a system. Situation Awareness itself is embedded in a larger framework which basically has the same properties as feedback control; see Figure 5.4.

Within this framework, situation awareness refers to the mental image that the regulator has of the system that he is controlling. Situation awareness is divided into three levels. The first

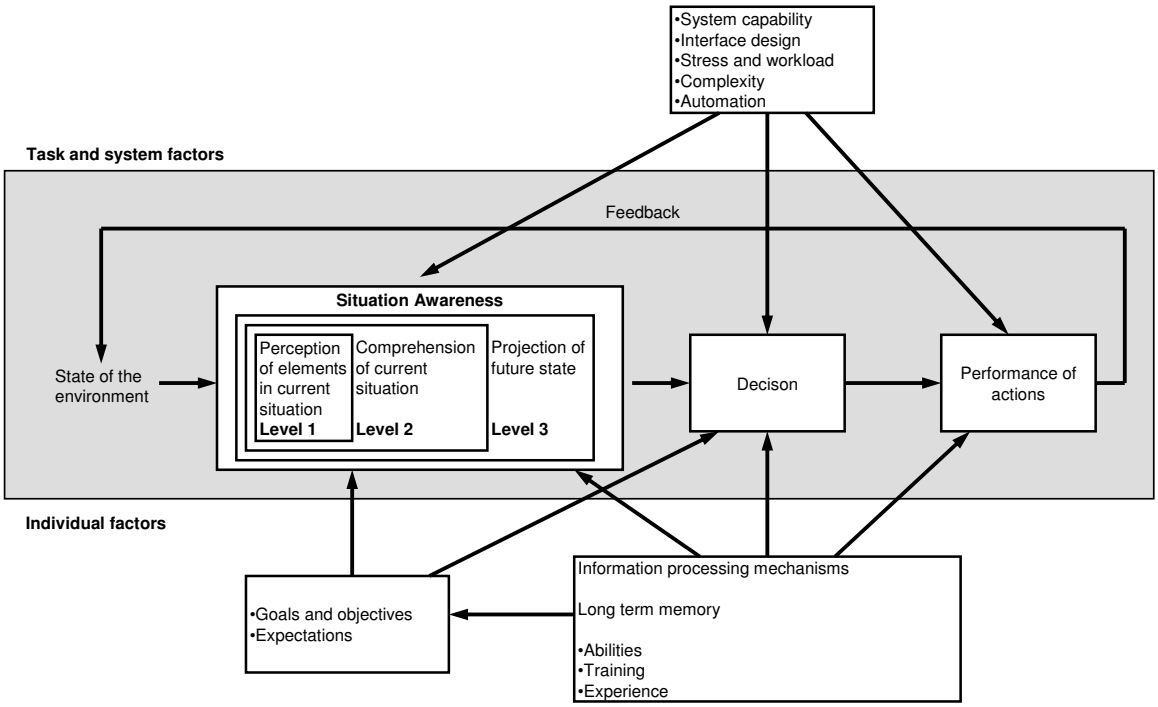


Figure 5.4: The Situation Awareness model (slightly adapted from Endsley, 1995).

level means that relevant elements of the environment have been perceived, the second level that their meaning is understood, and the third level that this is projected into the future. The mental picture that thus arises, serves to take a decision and perform some action. Situation awareness is one of the most important factors in decision making. Pilots who are good decision makers, have been shown to take longer in understanding a problem, but to select and execute the subsequent choice more rapidly (Wickens and Hollands 2000). The action that is performed changes the environment which closes the feedback loop. Broadly speaking, these are the same phases as in the DDA sequence, although certain steps of the diagnose phase are made somewhat more explicit. Also, situation awareness provides additional links to human's cognitive properties, such as incorporation of memory, expectancies and training. The DDA concept therefore allows us to makes links with this field of decision making.

5.4.4 DDA in decision making theory: bounded rationality

Limited abilities of humans to perceive and process information led to the concept of bounded rationality, which stems originally from economic decision theory (see e.g. Simon 1957). The concept is in contrast with the theory of rational decision making, where a decision maker has full information (detect), unlimited processing capacity (diagnose) and thus comes to an optimal choice (to be acted upon) in which all alternatives were considered. In reality, so it is argued, a human decision maker has only limited information and limited cognitive capacities. As a result, under bounded rationality, an individual only uses salient information, and searches until a satisfactory (instead of the optimal) solution has been found. The criterion which distinguishes between satisfactory and unsatisfactory solutions is called the 'aspiration level' and depends on conditions. If a solution appears difficult to reach, the aspiration level is lowered. The desirability of options thus changes over time; preferences are not constant (Sivak 2002; Elvik 2005). Especially in difficult tasks with many variables and decision rules, a human decision maker thus has to decide to take only a limited number of variables into account (the same mechanism is also one of the reasons that management information has to be filtered). Although one might expect that the quality of the solutions might decrease if they were found with bounded rationality, this is not necessarily the case, as has been shown by Gigerenzer and Goldstein (1996).

5.4.5 DDA in organisational learning

Every actor needs a model of the system in which he is working. By definition, a model is an incomplete representation of the reality which takes only those aspects into account that are considered relevant for the purpose of the model. Bounded rationality in fact represents the boundary of any (personal, mental) decision model. More experienced staff have wider bounds, which corresponds to having a richer model and thus the ability to make better decisions. Learning is required to build these richer models. Argyris and Schön (1996) provide detailed information about organisational learning. They indicate that 'operational surprises' – situations that have a different outcome than what was intended or expected – should be reported to and looked for by a so-called learning agency, which has the task to analyse discrepancies, analyse their causes and address relevant actors to generate possible improvements. Clearly, these three steps are also DDA loops. Argyris and Schön differentiate between two learning loops: single-loop learning has an emphasis on the detection and correction of errors within a given set of governing variables (e.g. KPIs) and is linked to incremental change. Double-loop learning involves also interrogating the governing variables themselves and may lead to radical changes such as the wholesale revision of systems, or alterations in strategy. When making use of the cascade model in Chapter 10 the issue of organisational learning will be approached using this DDA concept.

5.4.6 DDA in process control safety

The German standard VDI/VDE2180 deals with the topic of process control safety, and stems from the chemical industry. The standard distinguishes the following functions in safeguarding operations (Verein Deutscher Ingenieure 2007):

- Control system for normal operation: meant to steer and regulate the process as intended.
- A monitoring system is meant to give a notification if non-critical deviations occur, e.g. when the process is running out of the allowed tolerances. Although the quality will be lower than intended, the process could be continued from a safety viewpoint.
- Protection system: meant to prevent intolerable, safety-critical error conditions. A protection system generally checks whether one or more parameters are in accordance with permissible values. The functions of the protection system have priority over those of the basic process control systems. The activation of protection systems is very rare compared to the number of times the control and monitoring systems are used, because of their staggered configuration.
- Damage control systems: meant to limit the severity of damage once an undesired event has occurred.

Each of these control systems have to measure a variable value, diagnose it and decide which action is necessary (the normal DDA sequence), but they all are triggered on different variable values or different variables. An example could be an industrial oven in which bread is produced. A (thermostatic) control system would then regulate the temperature towards a value that is required to deliver high-quality bread. The monitoring system would provide an alarm if the temperature were to become too high or too low, to indicate that the bread is either not done yet or that it might become burnt. This is not yet unsafe (at least for the factory, though it could be unhealthy to eat), but it results in bread that cannot be sold. A protection system might shut down the oven once the temperature rises to a level which might result in fire. A fire alarm and fire doors can finally be considered to be damage control systems.

From the description of railway operation in Chapter 2, there are no indications that there are explicit tolerances defined for the process to be in, such as a time window within which a train path should be produced, so that there can be no real time monitoring system active. Indeed, rail traffic operation seems to have only a control system (the timetable and its delays)

and a protection system (the vital systems mentioned in section 2.3: the interlocking and automatic train protection system), but no real time monitoring system. It can, in that case, be expected that the process will relatively often have to be 'caught' by the vital protection systems.

5.4.7 Detect – Diagnose – Act in several fields

As the previous sections have shown, DDA is widely found as concept of control in several disciplines and is therefore a promising modelling concept for this thesis. It can be used on the level of individuals and of organisations, and make insights from many fields compatible. The next section will elaborate on the use of DDA cycles in the Eindhoven accident and show that the concept can also be used when several actors have to cooperate, as is the case in railways.

5.5 Discussion

5.5.1 DDA Examples from the Eindhoven accident

Table 5.1 will provide an example of the DDA sequences as they were performed by the driver of the local train in the Eindhoven accident. For a detailed description of the accident sequence, refer to section 4.6 and Figure 4.6, from which also the letters are used to indicate the subsequent phases in the accident. The information in the table is taken from the original accident investigation report (Spoorwegongevallenraad 1993).

Table 5.1: The Eindhoven SPAD in successive DDA sequences for the local train.

Step	Description of accident step	Detect	Diagnose	Act
A	Departure	Platform signal showing yellow	Depart + prepare to stop at next signal	Depart + remember necessity to stop whilst running from yellow towards red aspect (approx. 1 min)
B	Prepare to stop	Signal in 300 m distance in curve showing yellow (erroneous)	No further action required	n/a
C	Pass signal with red aspect	No detection because driver thought he had already seen the signal aspect	n/a	n/a
D	Run between red signal and danger point	No detection possible because no information was presented	n/a	n/a
E	Realise something is wrong	Points net set correctly	Something wrong, possibility of derailment	Apply emergency brakes
F	Emergency stop	(implicitly)	Driver understood that he stood in the way for the approaching intercity train	Tried to move backwards, but this was impossible since brakes were not yet released
K	Collision	-	-	-

The message of the yellow signal was somewhat ambiguous: it allowed the driver to accelerate, but simultaneously it also warned him to slow down so that the train could be stopped before the next signal. In the design of signalling systems, the driver is given approximately 9 seconds (see e.g. (IRSE 1995) for a comparison of European signalling systems) to detect the signal and read its aspect, and then he has to brake and stop before it. This means that the detection and diagnosis in the Eindhoven case must be made about 300 m before the red signal, in a bend, making it hard to distinguish which signal belongs to which track. This point 300 m in advance, where in fact a new DDA loop must be started, is not explicitly marked. As the driver was possibly confused by the mixed message about the meaning of the yellow aspect on departure, and the signals were located in a bend, the correct detection of the red signal failed. At this point, the DDA approach provides several clues for improved safety. For example, an additional message at 300 m before the red signal, indicating 'brake now!' could have helped the driver before step B. Similarly, a clearer relationship between track and signal could have been helpful to detect the correct signal.

When the driver approached the red signal, he actually detected a signal that was showing 'yellow' (the wrong signal, though) and interpreted it as permission for him to proceed. Since this sequence of events occurs frequently at this point in the system, i.e. the route ahead is often cleared in the meantime, the driver's expectation was possibly raised and became dominant, making the confusion of the yellow and the red signal in the bend of the track more likely. The driver thus diagnosed that no action was required, and as a result of this mental representation of the scenario, he no longer expected a signal with a 'stop' aspect, so he probably stopped looking for it and did not see it. Then, he detected that points were not set in the correct position, diagnosed that something must be wrong and he decided to stop and acted accordingly. He detected that he was in someone else's route, diagnosed that this was a dangerous situation, and he decided to drive the train a few meters backwards to clear the switch. However, since this action failed, it was not an effective control measure.

We can therefore conclude that performing a complete and correct DDA loop is a prerequisite for effective control in order to remain within the envelope's boundaries. This approach works to analyse accidents. But, if it is known upfront that a driver must perform DDA sequences, this knowledge can also be used to design a control system and its communication links. For example, the driver was not able to detect any information about the traffic controller's plan, since no such information was provided. The controller knew that the train would have to stop at the next red signal to let the fast train through, but did not communicate this to the train driver. Such information could have been helpful to guide the driver's expectation of what would happen. Similarly, also the traffic controller did not receive a warning that one of the trains under his control was behaving different than expected. Although he could have seen that a train was going through a red light, this is a very rare event which he would probably not have noticed.

5.5.2 Multiple actors

In any scenario in which an envelope's boundary is likely to be crossed, this must be clearly indicated to the controlling actors, so that they can subsequently detect this, diagnose the information and take action in order to stay out of the unsafe area. This, in fact, is the monitoring system's function as mentioned in 5.4.6. For instance, the driver must be shown how he should behave in order to drive the train according to the (re)scheduled wishes of the traffic control process or the timetabling process. The traffic controller must be shown which trains are not behaving conform his plan, and especially where and when a route conflict is upcoming, so that his attention can be steered towards those places where it is most needed and he can modify routes and assign them to trains in time. As we have seen in section 5.2.5,

this kind of information is only useful when it fits the mindset of the receiver, so that the operational envelopes must be geared to each other. This gearing can take place by exchanging information, preferably in variables that are meaningful in both actors' processes. However, currently the driver is told not to have a mindset, but to be a tabula rasa ready to obey signals which impose the controller's mindset. Chapters 7 and 8 of this thesis will show that this approach leads to problems, since it is impossible not to have a mindset. The only other mechanism available is to communicate much more intensely about each others mindsets, so that they synchronise.

Since a train driver often, especially at speed, cannot perceive the information that is relevant for his task unaided, he cannot start a DDA loop unless someone else or something else is supplying the relevant information. The traffic controller can detect useful information about the safe envelope of a train (regarding the goal state or the boundaries because of the position of other trains), but he cannot act directly on the train's behaviour. However, together they can perform a complete DDA sequence. In that case, the DDA loop of the traffic controller is complete when his action is 'inform driver'. The driver has to detect this incoming message and diagnose it, and take the required action. This can also be seen as a DDA loop that is extended over more than one actor, as shown in Figure 5.5. Because a traffic controller sets the routes for trains, a full control action in railways implies almost always this kind of cooperation between traffic controller and driver. Especially in distributed control systems such as railways, communication is therefore vitally important.

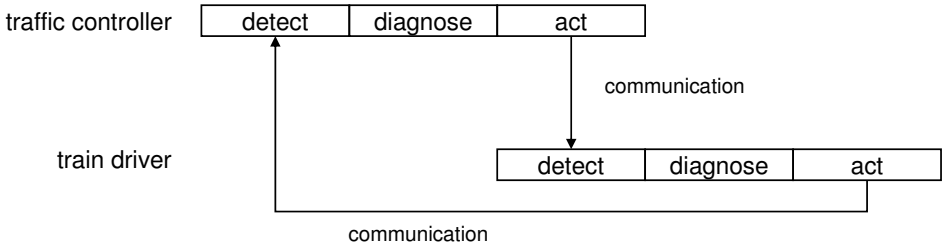


Figure 5.5: DDA loop extended over two actors.

As was already suggested by the cascade model, such control hierarchies make the goal and the boundary conditions at level n dependent on the action at level n+1. Thus, also the cascade model can be connected to cybernetics. Each function in the cascade model influences possibilities for consequent functions and thus sets new goals or operational envelopes for these. It is therefore necessary that these interacting control processes have coherent information flows about the time windows, the geographical restrictions and other limitations within which the processes should be completed. These are, in fact, all operational envelopes.

An important note is that the weaker the regulatory ability of a control loop, the more levels are necessary to produce the required amount of variety (Aulin-Ahmavaara 1979). However, 'the more levels the perception and action signals have to pass through, the more they are likely to suffer from noise, corruption or delays'. Therefore, if possible, it is best to maximise the regulatory ability of a single layer, and thus minimise the number of requisite layers (Heylighen and Joslyn 2001), so that it appears to be beneficial if the lower control levels can control their processes relatively autonomously. An explicit tolerance within which they should operate is an essential prerequisite to do so.

It is the task of designers and managers to empower operational staff to perform their tasks well. Hale et.al (2007) state that the function of safety management is to make sure that appropriate barriers (consisting of the three DDA parts) are in place and work effectively for

the whole of their lifetime. This means that the DDA concept is recursive, since management itself also requires to perform DDA cycles in order to know where it should place barriers or provide control measures to lower level actors. We see that infrastructure designers, traffic controllers and train drivers have DDA sequences that can be nested, as was already made plausible by the cascade model.

5.6 Conclusion

This chapter has given insight into what must be done to steer the process such that it remains within the boundaries of, for example, the safe envelope of operations. This chapter and the previous one have shown that the safe envelope and the detect-diagnose-act sequence can be combined to model process control with regard to both safety and quality. Table 5.2 summarises the similarities between cybernetics, which formed the basis for process control in this chapter, and the detect-diagnose-act sequence from safety science.

Table 5.2: Comparison of cybernetic requirements to the DDA sequence.

Safety science	Cybernetic entity	Cybernetic requirement	Meaning
Classical barrier	B	(passive measure)	Prevent the deviation from influencing the system
Detect	D→R or E→R	Information	Determine actual situation
Diagnose	R	Model	Extrapolating future situation without and with possible intervention(s)
Act	R→T	Control measures	Change situation
(unsafe states)	E	Goal	Gives direction for changes

The detect-diagnose-act sequence is generic and as was shown, it can be used to describe human behaviour and the functioning of technology. Well known concepts such as buffer times in railways or barriers in safety science can also be covered by this approach. It can also be used to model design and management: their functions are to make sure that control loops are in place, which itself can also be described as control loops. All functions in the control model (e.g. cascade model) need to detect where the envelope's boundaries are, to diagnose the situation and to perform actions. One possible action can be to provide a buffer or barrier to a lower level (which serves as a pre-defined action, i.e. open loop control) or to provide tools or information to enhance lower level DDA sequences.

There are several ways to cope with disturbances, such as:
- Eliminate them (which is possible in some cases, but hardly ever in all of them);
- Increase buffering capacity to block their consequences (costs capacity);
- Lower the standards of what is acceptable (at the cost of quality and/or safety);
- Make them controllable (costs effort).

It can easily be seen that improved control is a prerequisite for a better utilised railway network, since the other three options will not lead to better utilisation. Effective control requires that the intended controller of the process can detect relevant information, diagnose it, and influence the process. Since effective control requires that control measures are taken in due time, it is preferable, if not essential, that regulators on any level have the ability to

anticipate; also, it is preferable to keep the number of hierarchically placed regulators in a control system as low as possible, to avoid noise and delay due to additional communication.

The previous chapters introduced knowledge from many different fields. The next chapter will summarise the mix of models and methods of which the theoretical framework consists. This framework of nested control loops will serve as a reference for the Chapters 7 to 9, where several control functions are analysed.

6 The theoretical framework: a mix of models and methods

6.1 Introduction

Chapters 4 and 5 have shown that it is important to know how the process of supplying train paths is being controlled. To do so, several models were introduced: the cascade model, the safe envelope of operations, detect-diagnose-act (DDA) sequences and situation awareness. These will be combined in section 2 of this chapter, where also the validity of the framework will be discussed.

In the next section (6.3), it will be explained how data were gathered and processed to analyse the railway's control processes. A dual approach was taken: data were gathered and analysed using a grounded theory approach, for which many interviews have been conducted and their results were combined with data from literature. This section will indicate:

- what grounded theory is;
- how the interviews were set up;
- how the conversations with the interviewees were structured;
- how the data were processed.

Besides the grounded theory approach, data were analysed using the theoretical framework that was developed in the first part of this thesis.

The following three chapters present the results of these analyses. The processes of train control and traffic control, as the key elements in this dissertation, both deserve a separate chapter, Chapters 7 and 8 respectively. Chapter 9 will focus more on the functioning of the system as a whole and handles topics like influences from design and management; it has a more integrative role. It combines the outcomes of Chapters 7 and 8, the cascade model and organisational learning so as to allow for smooth cooperation of all required system functions.

6.2 Combining the models

6.2.1 Combining the safe envelope model, DDA loops and the cascade model

In the remainder of this thesis, several of the aforementioned models will repeatedly be used to analyse phenomena. To avoid confusion about the multitude of models, this section will explain how these models have a logical relationship:

- As was shown in Chapter 5, the right action to be taken notably by train controllers and drivers requires *situation awareness*: a mental representation of the system state. This representation is formed during the diagnosis phase of a DDA loop, so that the term situation awareness itself will not be used frequently during this thesis, but the concept is incorporated in DDA.
- The *safe envelope of operations* can help to specify which information is essential for operators to know: to achieve their goals, they must know the goal and the allowed tolerances, which we called operational envelopes. To remain safe whilst trying to reach their goal, they must be able to know where the viable envelope is and at which moment or place action ultimately must be taken in order to keep the system in a safe state (controllability envelope).
- Perceiving and diagnosing information items from the safe envelope of operations in turn requires correctly functioning *DDA loops* (detect – diagnose – act).
- Since much information stems from other functions, these DDA loops should be coherent throughout the cascade model. Feedback about mismatches between intended and actual outcome of an organisation's process, understanding why this is the case and then modifying processes such that the problem is solved, is organisational learning. The cascade model shows where in the system the decisions are made which successively determine the operational envelopes of lower functions and where feedback needs to be directed to, to ensure these decisions are optimised and take as full account as possible of the vagaries of real situations.

6.2.2 Validation of the theoretical framework

A first way to validate the framework is to make sure that it was constructed properly by regularly checking for inconsistencies in the construction process and by showing the result to experts. These checks were performed when the safe envelope of operations (roughly corresponding to Chapter 4) was sent to two experts; they did indicate that the model would lead at least to similar findings as a range of already validated existing accident models, which is an indication for the model's concurrent validity. Similarly, the review of this thesis by the examination committee serves as such a check on validity.

The risk of developing a model from any given viewpoint is that it comes with advice that is incompatible with or even contradicting recommendations formulated from other viewpoints (Veeneman 2002). The previous chapter showed that several fields (human factors, management science, and control theory) can be described by the same approach with DDA sequences, even though they look different at a first glance. This means that the same theoretical framework is applicable in different fields, which reduces the probability of coming up with conclusions and explanations that do not fit those of other fields, and provides a certain face validity for this approach.

The best way to validate a model is to compare its predictions to reality. This requires a dataset that is independent from the data that were used to construct the model. Unfortunately, this type of validation cannot be performed at this stage, since no such second dataset exists yet. In the next three chapters, such data will be collected and analysed using the framework. This can be considered as a first validation of its ability to analyse data from the past

(interviews and document analysis) and diagnose gaps in past projects. The model under consideration is mainly meant, at this stage, to help understand mechanisms, not to make (numerical) predictions. A next test, outside this thesis, could be to use the framework to design new projects and to predict that they will work if designed according to the model.

The first application of the framework in the following chapters may also result in possible improvements to the framework itself. That application will be critically reviewed in the coming chapters to see whether such improvements are possible. An improved framework would facilitate even more understanding, and understanding is required before effective solutions can be developed. Therefore, Chapter 10 will start with a review of the framework used and the application so far. On the basis of this suggestions will be given for improvements.

The next step in this research, and all chapters between this point and Chapter 10, is to use the framework to analyse data from actual operations.

6.3 Collecting and processing of data using a grounded theory approach

6.3.1 The concept of grounded theory

Two ways of data processing
Besides the safe envelope framework, a grounded theory approach is used in this research project to collect data and find explanations for the phenomena we do not fully understand yet. In a grounded theory approach, the data convey their own messages resulting in so-called emergent categories. The emergent categories were formed first of all, at least in a conceptual way, because to let the data speak for themselves, it is necessary that researcher is as unbiased as possible. It is therefore important that these categories are formed before the data is used with the safe envelope framework. The approach will be described in this section and the final results will be presented in section 9.6.

Once the emergent categories were roughly formed, the same data could also be analysed using the safe envelope framework as described in the previous section. These analyses are made at the end of the Chapters 7 to 9.

This means that the data collected during this research were processed in two ways. For ease of understanding by the reader, the order of the coming chapters is such that findings from either one approach are grouped according to their function (e.g. train driving, traffic control, systems design). The order in which you will read the next three chapters is therefore not always the same order in which the information was derived.

From raw data to a theory
Grounded theory is a method to construct hypotheses from observations. It was developed by Barney Glaser and Anselm Strauss (1967). It is an inductive method, as is every method to seek general explanations for observed phenomena. Its key notion is to let data speak for themselves instead of pushing them into an existing framework – it is the framework that is to be derived from data, not the other way around. Grounded theory helps to get rid of blurry details, which are common to situation-specific cases, and step by step to come to a theory containing the key elements relevant to an (hypothetical) explanation for the phenomena observed. The hypotheses that are constructed in this way can be tested in a later stadium.

Multiple data sources
In order to gain a rich understanding of the research topic and thus to develop a reliable explanation for the phenomena observed, it is necessary to develop a theory from multiple disciplines in order to be reliable (Haig 1995). In this study the following information sources were used:
- observing of operational staff during their work
- interviews with operational staff
- interviews with subject matter experts
- scientific and specialist literature on a variety of topics (human factors, safety science, railway technology, timetabling, signalling, etc.)

6.3.2 Preparation of the interviews

Selection of the interviewees: maximum variety
The group of interviewees has been sampled purposively in order to collect as much coherent data as possible from as many viewpoints as possible. As the focus of the research is on operations, the main group of interviewees consisted of train drivers and traffic controllers. As traffic controllers have a central role in controlling the traffic process, we expected them to be able to reveal the most relevant topics and decided to start by interviewing them. The group of traffic controllers consisted of ten interviewees from five control centres. This avoided the interview results being biased by very specific, location-dependent issues, and enabled me at the same time to get a broad view of these local issues and compare them to each other. For the same reason, the group of train drivers consisted of seven passenger train drivers and three freight train drivers, from a total of four train operating companies. The group was extended as data collection (by interviews, but also by literature research) continued. Interviews were added as long as it was expected from other data that a certain interview could yield useful information. For example, some traffic controllers described certain phenomena that are known in literature as the out-of-the-loop performance problem (Endsley and Kiris 1995), which is usually caused by flaws in system design. This was a reason to add an interview with designers of the corresponding system and to consult relevant literature. In grounded theory there is no attempt to collect data from a representative sample of the total population. The concentration is on richness and variety, so interviews are continued until they deliver no significant new insights. This way of sampling gives a broad view on staff's perceptions of problems, probable causes, and ways in which staff deal with them and also their ideas about possible improvements – instead of some 'average problem' with 'average solutions'. In the end, the following people were interviewed in sessions that took approximately two hours per interviewee:
- Drivers (10, from four companies)
- Traffic controllers (10, on five control centres)
- Regional traffic controllers (4; on two control centres)
- Timetable designers (2, one from ProRail, one from a train operating company)
- Infrastructure designers (2, one from ProRail, one from a train operating company)
- System designers (2 for signalling and 1 for the VPT traffic control system)

Preparation of a questionnaire
The safe envelope framework served as a basis for preparing a questionnaire. Because the interviews had to be conducted in an early stage of the research, a preliminary version of the safe envelope model was used, in which the DDA-cycles did not yet have the prominent place that they have acquired later, as was explained in the previous chapters of this thesis. For the resulting questionnaire this was not a serious problem as the essential elements of the model had been developed and control-related questions were included because this was a central

element in the research question. The questionnaire contained questions about the following (partly overlapping) topics:

- *Goals* of the interviewee, as a warming up question to let the interviewee think of what he wants to achieve.
- *Current practice*, problems therein and workarounds, with regard to the operational envelope, controllability envelope and viable envelope. These boundaries were not explicitly mentioned during the interviews, but served to develop more detailed questions about each of these. The preliminary version of the safe envelope framework did not yet contain the attainability boundary, but some interviewees implicitly already referred to it in the sense that they sometimes could not reach a desired solution in time.
- *Information needed* in order to control the process such that it reaches the goal and does not exceed boundaries.
- Ideas on *future improvements*; to make use of possible practical ideas on this point and prevent re-invention of the wheel.
- *Collaboration with other staff members* in order to identify the cooperation in the cascade model and especially to check the existence of feedback loops; under this heading, also explanations for SPADs were asked for.
- Any *other items*? And which of the above mentioned problems are most important? Other items were generally not mentioned, which indicates that the questionnaire covered the topics that were considered relevant by the interviewees.

The questionnaire (in Dutch) for train drivers can be found in Appendix B. The questions for local traffic controllers and regional traffic controllers were the same, although changes were made with regard to the cooperating actors as relevant. For the subject matter experts, either the same list was used and irrelevant questions skipped, or specific questions were asked depending on the purpose of the interview. The questionnaire was sent to the interviewee in advance so that he/she could prepare for the interview, although it turned out that only some of them indeed did so. In all other cases, the purpose of the interview was briefly explained before the interview actually started.

6.3.3 Performing the interviews

Anonymous processing

It was expected that many interviewees would not always work exactly in the way the rules prescribe, thus making their tasks easier and/or getting things done where the process would otherwise stall. Since rule breaking is frowned upon in railways, it was believed necessary to reassure respondents about confidentiality to assist them in admitting to it. Therefore, all interviewees were told that the data they would bring forward would be processed anonymously, so that they could speak openly about how they do their work – even if this was not in conformance with the rules. It appeared that none of the interviewees had hesitations to do so.

Some interviews with assistance

The series of interviews with traffic controllers were conducted by me, as interviewer, with the assistance of a former employee of ProRail's traffic control department, who retired some months before the interviews took place. This had a number of advantages:

- The tasks of taking notes could be handed over to the former ProRail employee so that there was no need to stop the interview to write down what was said. It was decided not to tape record the interview as this might frighten people from speaking openly. Later, when I was more acquainted with both the questionnaire and the subject itself, it turned out that the interviews could also be taken without further assistance.

- The former ProRail employee had decades of knowledge about the systems in use, which was useful to interpret some of the answers given, especially by the older interviewees.
- As he was just retired, his knowledge was still up to date but his presence did not hamper interviewees to speak openly.

Semi-structured interviews
The interviews were semi-structured. It depended partly on the interviewee how much emphasis was placed on each of the topics in the questionnaire. Also, the interview structure left space for the interviewee to bring forward other points of importance. Basically the interviewees could tell their story in the order they liked (sometimes jumping from one subject to another) and from time to time the interviewer checked which of the questions had been answered and steered towards topics that had not received so much attention.

6.3.4 Processing interview data

No statistical significance by design
Statistical significance is not meaningful in these types of interviews. Firstly because the group of interviewees was intentionally not chosen to be a statistically representative sample but one with maximum variety, so that most could be learned; secondly because statistical significance only says that the answers cannot be explained by *chance*, but it does not mean that the message of the majority is *true.* Let us assume that we ask a hundred traffic controllers whether they can rely on the interlocking to maintain the desired safety level. As they are trained that this is the case, probably the vast majority will say 'yes'. However, a small minority may have experienced that in certain conditions it is wise to do some extra checks, such as informing the train driver additionally. They will say 'no.' In this case, the argument that the interlocking is sufficient for safe operations could be statistically significant. However, this only means that this argument cannot be explained by chance (since it is explained by the education of the traffic controllers) but the explanation for SPAD's could still be found in the experiences of the minority. It is not the statistics that count, but the quality of reasoning and the underpinning of that with arguments.

Interpretation of interview results: forming of categories
Qualitative research is often inductive as it focuses on understanding phenomena and building theories about the mechanisms causing these phenomena – so it works inductively from a single case or a few cases to a general theory. This is a process that involves thinking and because thinking is a process that can be learned and steered only to a certain extent (Boeije 2005), this is a subjective part in doing research. This is also recognised by Strauss and Corbin (1998), yet they indicate that *"theory derived from [acquired] data is more likely to resemble 'reality' than is theory derived by putting together a series of concepts based on experience [of the researcher] or solely thorough speculation (how one thinks things ought to work)"*. In order to stay as objective as possible it is important to compare the theory as it is formed to the data, to be sceptical and to look for counterevidence.

By analysing, interpreting and combining data and by comparing that to literature, certain categories of data will form. This is the basis of the construction of a 'grounded theory', a set of explanations why certain phenomena exist and a way to structure the thinking process by which the theory is developed. Finding those theories is a process in which data are reduced: from detailed data in different sources of information, via certain categories of data, more abstract concepts can be derived. The more abstract a theory is, the more broadly applicable it is, but the less it will be grounded in data (Strauss and Corbin 1998). All this is an iterative process: it can very well be the case that certain phenomena in the data have not been

identified, because the emerging theory was not yet detailed enough. For the researcher, things were at that moment not apparent because it was not clear what to look for in the data. A researcher 'must have a perspective (to) see relevant data and abstract significant categories from (it)' (Glaser and Strauss 1967). Once the theories evolve further, it is possible to see more and other explanations for phenomena that could not be seen earlier. The data help to develop the theory and vice versa. When analysis of additional data does not lead to new insights, a category is 'saturated' and data collection can be stopped (until for some reason additional data appears to be necessary to answer a question that may arise somewhere else). Forming of categories was done by creating a large table in which the types of interviewees (corresponding to the functions of the cascade model) were set out along one axis and the interview topics along the other, so that statements of interviewees could be sorted into these cells. Statements of a similar nature were all given a colour, which was the beginning of an emerging category. An example was where interviewees stated that they were discouraged. Every statement could have multiple colours. Statements without a colour that did not fit any

Table 6.1: Grid for the forming of a grounded theory with quotes as examples.

	System designers	Infrastructure designer	Timetable designer	Regional traffic controller	Local traffic controller	Driver
What is your goal?		Enable trains to run			Having a happy customer	Move the train / the freight
Who is your customer?				TOC / passenger		Passenger / shipper
System designers					The VPT system was meant to…	
Infrastructure designer						Signals are not clear in case x and y
Timetable designer						How timetable design helps the driver
Regional traffic controller						
Local traffic controller						
Driver		How drivers help infra-structure design				The driving task
Whole chain				Don't know to whom address feedback		
SPAD causes					Higher frequencies	Shorter education
Management influences						KPIs not in line with real goals
Ideas for improvement				increase predictability		

of the existing categories either indicated that another category should be formed, or that the statement should be discarded as being irrelevant. Once categories started to emerge, the table was analysed again to see whether some of the categories that were found at a late stage were also underpinned by statement that were analysed earlier. In this way, the categories were slowly saturated.

The table in its full extent consisted of six pages on A3 paper, so it is not possible to include these intermediate results in this thesis (at least not in an intelligible way), but the structure of the table is represented in Table 6.1. The example text in the cells shows how data were sorted in the table. In the six page version nearly all cells were filled.

The resulting categories will be shown in section 9.6. After the forming of categories, the interview data were also checked against the (now fully developed) framework of the safe envelope and the DDA model, and compared to literature.

6.3.5 Verification and validation of the grounded theory

Since generating a theory is an inductive process by definition, formal validation cannot be done; it can only be invalidated when hypotheses, that were derived from the theory, are later tested (Christensen 2006). Nevertheless, several steps can be and were taken in order to make sure that the grounded theory made sense and to check that the theory-in-progress was still grounded in the interview data and not accidentally drifting apart from that. In grounded theory this process is generally called validation (Strauss and Corbin 1998), although other sources on methodology would call this process verification (de Groot 1961).
In order to verify the factors found, it is necessary to:

- Make sure that the theory does not drift apart from original interview data,
- look for counterevidence continuously, and
- a 'final external check on trustworthiness can be facilitated wherever practical by presenting derived concepts and models to significant independent individuals' (Pidgeon et al. 1991).

All of these steps were taken consistently through this study.

The series of interviews were planned with sufficient time in between for transcription and analysis, so that they could be processed when they were still fresh in the mind. After the interview the notes were ordered by the interviewer and the transcriptions were then checked by the interviewees, which is critical in order to allow them to check the initial interpretations. This procedure is in line with that recommended by Pidgeon et.al. in their paper about using grounded theory in an engineering context (Pidgeon et al. 1991). For the groups of drivers, local traffic controllers and regional traffic controllers, results have been combined into one report for each group, indicating the key elements on which the group agreed and those elements on which answers varied considerably. To check the correctness of these reports, these were again circulated to all interviewees and additionally to ProRail's staff members to allow for feedback. In the case of the traffic control interview reports, the reports were also sent to traffic control staff. As far as feedback was received, this yielded positive results. However, especially from the driver reports, not many responses were received; a possible difficulty with the method at this point may be that drivers are not so much interested in reading reports, especially not when it took them also time to check the initial interview transcriptions. Note that most drivers did so in their spare time. Nonetheless, later conversations with drivers based on the growing theory did not lead to indications that the theory was in need of modifications.

In a further stadium, the interview results were also compared to literature from several relevant fields, such as safety science, railway operations engineering, human factors, and management science; thus, it was possible to see whether or not the reported phenomena must be seen as a problem or not. A final check on trustworthiness has been performed on relevant parts of this thesis by a former train driver who is now an independent management consultant for a number of railway companies in several countries, and also by a team of specialists from several ProRail departments who commented on this thesis.

6.4 Conclusion

This chapter discussed how the several models and methods used during this research project fit together. The cascade model, the DDA sequences, the safe envelope of operations and the situation awareness model can be used together to describe which information the controller of a system needs, where this information can come from, and what must be done in order to process the information successfully.

The second part of the chapter dealt with the question of how explanations for SPADs and quality of service problems can be derived from data. When certain phenomena are observed in real life, they are generally obscured by all kinds of fuzzy details and therefore the phenomena cannot be understood easily. Grounded theory is a method to abstract key explanatory factors from a number of cases, thus getting rid of the case-specific details and leading to a more or less reliable theory that is grounded in the data that have been used to construct the theory. This theory will unfold itself in the next three chapters.

PART III: Understanding the current situation

7 The train control process

7.1 Introduction

Chapters 4 and 5 have shown that it is important to know how the process of operating railways is being controlled in order to provide transport services safely and efficiently. In the quest to understand how the occurrence of delays and SPADs can be explained, this chapter and the next deal with two key processes which try to control these phenomena. They correspond to the two lowest levels in the cascade model:

- the driver's process: controlling a train's speed on a given route (this chapter);
- the traffic management process: rescheduling trains and providing routes for trains when they need them (Chapter 8).

Figure 7.1 summarises the cooperation between traffic controller and driver.

Although it was assumed that these control processes are well understood, it turned out during the research that the formal process descriptions available in the companies do not fully cover what is happening in practice. Additional analysis was undertaken to find out which quality and safety problems operators see themselves confronted with and how they solve them. In order to understand the information in this chapter, it is presumed that readers have made themselves familiar with the background knowledge provided in Chapter 2.

The remainder of this chapter is structured as follows:

- this chapter starts with an analysis of signalling from a control perspective in section 7.2 using the DDA model described in Chapter 5;
- in section 7.3 the analysis will be continued with a focus on the Dutch signalling system NS '55;
- in section 7.4 the human factors issues such as vigilance, route knowledge, driver expectancies and distraction will be discussed;
- section 7.5 will describe the findings in terms of the safe envelope model;
- in section 7.6 some conclusions will be drawn.

The information in this chapter comes from the literature, discussions with those concerned at all levels in the rail system, but above all from the interviews with train drivers (and to a

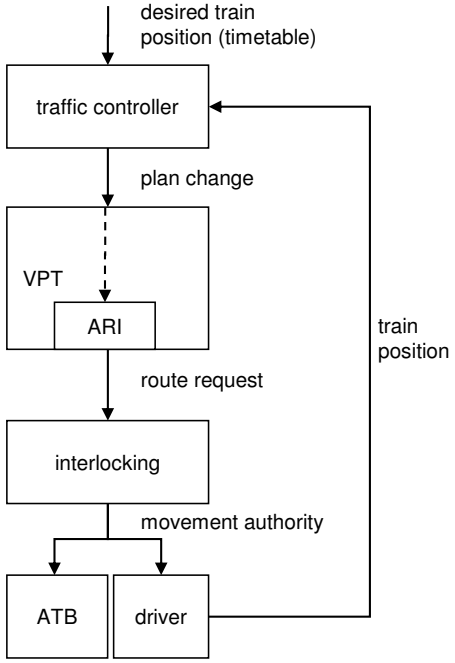


Figure 7.1: Interaction between traffic controller, control system, interlocking and drivers.

lesser extent with traffic controllers) conducted for this thesis; see Chapter 6 on the methodology. We will see that the driver requires more, but also more precise information to form a correct picture of the environment in which he has to act. This requires support by traffic control and designers of driver support systems.

7.2 Signalling from a control perspective

7.2.1 Goal and boundary conditions for train control

Realisation of desired system states: the timetable as reference model
The function of a railway system is to provide quick, cheap and predictable transportation of passengers and goods between a large number of locations. For the users of train services, the predictability of this mode of transportation is provided by the railway timetable. Another important function of the timetable is to provide a reference model for the organisation and its control of the railway system. Strangely enough, this use of the timetable is not well developed. The timetable is constructed of a large number of intended trajectories (lines relating train positions to time) organised so as to avoid conflicts between trains (i.e. lines being too close to each other in time and space such that blocking times would overlap). The absence of conflicts in the timetable means that, if the schedule is executed properly, a train departing a certain location will reach its next destination without encountering any conflicts, provided the train adheres to the planned speed, which is the first derivative of the timeline.

One would expect that the train control system would make optimal use of this conflict-free reference model, encouraging close adherence to the planned timelines. This, however, is not the case. The current signalling system is not set up to control a train's speed according to the prescribed timeline, but mainly to handle foreseen and unforeseen conflicts safely. Controlling the train's timeline is left to the train driver, who generally is not informed about changes in the plan, and cannot see in sufficient detail if he is adhering to the plan, and therefore exercises a form of open-loop control. In relatively low usage networks, this rather primitive form of control actually works, since the builders of the timetable include buffer time (unused time between timelines), allowing for limited deviations from the timelines. If

the buffer times provided in the timetable are not sufficient to handle timetable deviations, a conflict between two movements will result. With increasing network usage, as will be the case in the Netherlands, buffer times tend to decrease, which may lead to a rapid increase in the number of unplanned conflicts. The signalling system will then be relied on more frequently and have to prove its worth.

Signalling background knowledge
Signalling is one of the key elements to rail operations, since it is one of the factors that make rail transport the safest mode of transport and it is also unique to railways. However, not so much has been written about signalling. Most European railway companies have developed their own system. As a consequence, knowledge about most systems currently in use was written in the native language and remained within the company – if it was written down at all (Pachl 2008). One infrastructure designer, who was generally referred to as the expert with respect to the Dutch signalling system NS'55, indicated that approximately 50% of the knowledge of the signalling system was never written down and is only present in the minds of, generally older, people. Similar concerns were also raised by one of the senior experts of ProRail's train protection department, who said that, since all projects have deadlines, these do not allow for gathering knowledge to understand enough of the general system into which they will fit; this depends totally on the individual's initiative.

From a number of interviews, it became clear that the Dutch signalling system is considered to be one of the most user-friendly lineside signalling systems in the world, as its information can be easily processed by drivers. Middelraad (2000) gives an excellent and complete description of the system's realisation in the 1950s, the design choices that were made, and the modifications that were made during its lifetime and the reasons for changes. In a book by the Institution of Railway Signal Engineers a number of signalling systems are compared to each other, amongst which the Dutch one (IRSE 1995). However, the systems are compared mostly on detailed technical aspects, (e.g. comparing the meaning of a flashing green light in several systems) but these are not related to the task of the train driver – which is somewhat odd, since signalling systems are in place to supply the driver with information; a more useful comparison would have been to list all possible messages that drivers can be told with different systems. Other authors focus on the human factors involved, of which Wilson et. al. (2007a) provide an excellent overview. However, the available literature does not yield an analysis of a line-side signalling system from a control perspective. This is therefore the topic of the next few sections.

Prevention of unsafe states: the function of signalling
NS, the former Dutch Railways, gave the following definition for the function and the goal of signalling in 1963 (AV130, ProRail 2005a):
With signalling, we mean giving instructions to the driver with regard to the safe and efficient flow of railway traffic. The goal of signalling is:

1. *To ensure safe operation;*
2. *To enable a smooth flow of traffic by preventing trains having to stop or slow down unnecessarily.*

In this definition, it is not clear how the second part of this goal should be interpreted. Taken literally, it refers to a traffic management function (such as providing advisory speeds) which cannot be performed by the signalling system. A more likely interpretation is that the signalling system's messages and the location where they are given should not unnecessarily obstruct smooth operations. But then, that is not a goal of signalling but a design requirement.

7.2.2 Essential information to control a train movement

Three essential variables
The essential variables to control a train movement are the train's position, its speed and its timing (adherence to timetable). The operational envelopes for these variables are:
- Timing: the difference from the reference time in the (actual) timetable should be such that the timetable remains feasible (with regard to the internal function of the timetable as a production schedule) and such that customer expectations are not violated (with regard to the external function as a service catalogue).
- Route (position): all track sections assigned to that specific train by the traffic controller are called a route for that train.
- Speed: all speeds between 0 km/h and the maximum speed, which is the lowest of the maximum speeds for either the track section or the rolling stock type.

The combination of speed and position in a speed-distance profile is often called the 'Movement Authority'. Normally, a movement authority is given via written orders, trackside signalling or cab signalling. In some cases, a driver is instructed to drive on sight, which means that he has to find one or more variable values himself. In all cases, a driver must be able to construct a representation (situational awareness) of the world on the basis of the information he receives. Ideally, his representation consists of the actual values of all the essential variables and their operational ranges, as was defined in section 4.9. However, in practice some parts of the information are missing, as we will see in the remainder of this chapter. Exceeding the movement authority (violating speed or position requirements) may directly impose a safety risk. Timing of trains has the closest relation to realising the intended timetable, but it can only be influenced indirectly via a change in the train's speed, for which having a movement authority is a necessary prerequisite.

Information about timing
As we have seen, drivers hardly have any information about their timing. They have information on the published timetable, but this timetable is rounded to whole minutes and not updated to the current, possibly disturbed, situation on the network. Since no actual control information is available, the timing requirements will tend not to be met, which means that trains are running early or late. Although such violations do not directly impose a safety risk, they may do so indirectly, because the production plan may become infeasible. The planned timetable contains time buffers (both in the running times of one train and between two trains), meant both to make the schedule fit and to allow for small disturbances to be absorbed. Train drivers, however, do not have information about the exact amount of time supplement that is contained in the schedule and where on the route this is planned to be, nor do they have detailed information on whether they are adhering to their planned paths. A freight train driver indicated that this leads to deviations from the intended timing of 2 minutes plus or minus. A passenger train driver gave an example: he departed from a station with a delay of a few minutes, but 20 minutes later he arrived at a red signal which could not be cleared yet because he was now, without knowing it, running ahead of schedule. Furthermore, differences in driver knowledge and driver preferences result in drivers choosing different driving strategies. Some drivers will run at maximum speed, for instance to save time for any unforeseen events. Others know where the system has play in it and adopt a somewhat slower driving style, to save energy and have a more comfortable ride. Time supplements, if used incorrectly, cause disturbances rather than absorbing them (Weeda 2005; Weeda and Hofstra 2008). However, drivers have no information on how to use the available running time in such a way that they can sufficiently adhere to the plan. If drivers know about supplements, this is by experience which will vary from driver to driver. Their knowledge

will vanish when a new timetable comes into force, and will become inappropriate when the actual situation is different from the normal situation.

Conflicts that might appear to be solved during the timetabling phase by adding extra running time may therefore still arise during operations. The conflict-free schedule is not really conflict-free. In reaction to conflicts, traffic controllers may update the plan by reordering trains in time or in space, to avoid unnecessary braking (which again leads to delays and energy losses). However, drivers complain that they are usually not informed about such changes. Their behaviour then cannot match the updated plan and drivers become only aware of this when faced with an unexpected signal aspect, which means that some of the conflicts remain. Performance analysts at ProRail, who compare the actual track use to the planned use, confirm that this type of conflict occurs (Schaafsma and Weeda 2009). Although no precise data are available, they estimated that conflicts remain and lead to an unplanned stop in the following percentages of cases:

- if planned conform minimal planning norms (containing approximately 1 min. buffer time): 5 – 10% of all cases;
- if planned without buffer time: 30% of all cases;
- if the first train is a local train, it is generally late because running time norms are too optimistic: the second train must stop in nearly 100% of all cases;
- if the second train is an intercity service with additional running time because it cannot enter the station earlier: it will arrive early and has to stop in nearly 100% of all cases.

Also, one would expect that if trains are scheduled for exactly the same time, they would conflict nearly always, such that half of the trains would be required to stop. In a trial to enhance capacity in the Schiphol tunnel this was tried out. However, as trains did not closely follow their scheduled paths, this happened to 'only' 30% of the trains (Projectteam Samensporen 2006).

When asked which additional information they would like to have, many drivers indicate information that in some form relates to the actual plan that the traffic controller has, such as:

- whether their schedule has been changed (to be able to drive more comfortably and more energy-efficiently, to prevent unnecessary braking);
- getting an advisory speed, either directly from the traffic controller or automated (especially for freight train drivers);
- the track to which the train will be sent (to determine where to locate the next signal);
- which train is running in front of them, so that they can drive more anticipatorily.

Similar results have been found in an extensive Swedish study (Kecklund et al. 2001).

It is clear that not rescheduling the plan or not telling drivers about changes will lead to timing errors, which can result in a conflict. The signalling system does not anticipate such future conflicts; it only performs a check on whether the safety conditions are met *at this moment.* As a result, the signalling system functions as a kind of trap when a train approaches an expected conflict. The train will not be slowed down until there is no other option for the driver than braking at the braking distance from the conflict point and eventually stopping. This of course is only true under the condition that the driver is not otherwise informed (by experience or by the traffic controller) that he should slow down to prevent reaching the conflict point too early.

Information about the route

Certain infrastructure elements, such as points or moveable bridges, have to be set to the correct position before the train can run. In order to drive safely, a number of infrastructure-related safety conditions must be met (AV101, ProRail 2005a):

- the points, bridges (etc.) in the route must be in the correct position,
- they should be locked in that position,
- the route should not be occupied by another train,
- none of the elements in the route should already be assigned to another route, and
- level crossings with road traffic must be closed in due time.

It is a task of the traffic controller to make sure that routes are set in a timely way and that they respect these conditions. Generally, drivers cannot check these conditions visually, as it would require the driver to see ahead *reliably* beyond the braking distance. The normative braking distances vary from 400 m at 40 km/h to 1150 m at 160 km/h (*Regeling Spoorverkeer* 2004). Even a distance of 400 m, required at the lowest speed of 40 km/h, cannot always be overseen in the case of poor weather conditions (e.g. fog, darkness) and on curved tracks, especially if the sight lines through curves are obstructed, e.g. by vegetation, constructions, or a train on a neighbouring track. Some form of signalling must be used to inform the driver about the situation ahead. Also, drivers are not informed about the track they will be switched to, so that they cannot know which signal is relevant for them, unless the station layout is relatively simple.

Information about speed

Speed is the variable linking time and position. It is also the variable defining how much kinetic energy a train carries, which in turn influences the required braking distance. The amount of energy in a system also determines the amount of possible damage if the energy is released in an uncontrolled way (Johnson 1973). The allowed *speed* at each location along the *route* can be plotted in an distance-speed diagram (Figure 7.2). This is called in some literature, a speed profile. Variants of this figure will be used throughout this chapter. Given the previous two variables, we can define three boundaries, each describing a typical situation about which the driver has to be informed:

A. There is a route for which no other restrictions apply than the maximum line speed and the maximum rolling stock speed (a horizontal limitation in Figure 7.2);
B. There is a route with restricted speed, e.g. in curves or on diverting points (a horizontal limitation in Figure 7.2);
C. The route will end and the train must be stopped before a certain point (a vertical limitation in Figure 7.2).

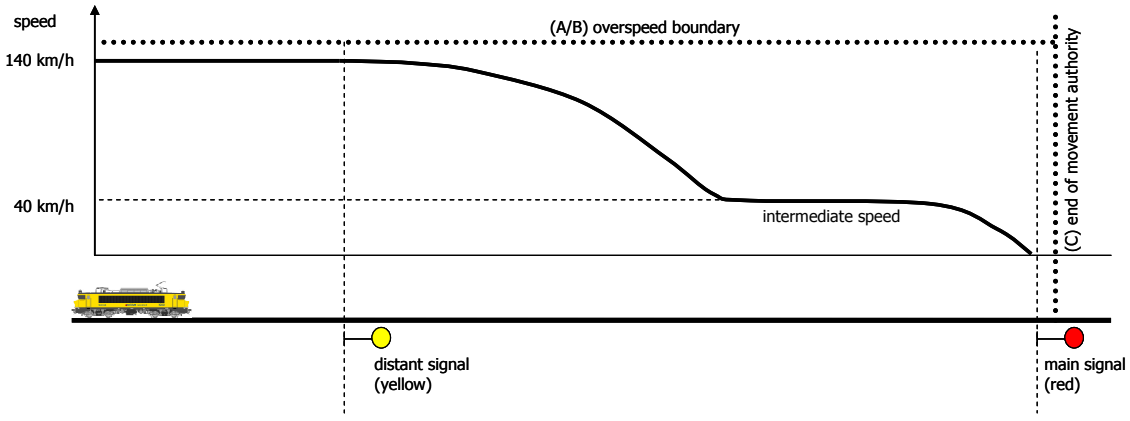


Figure 7.2: Speed-distance profile with characteristic, mandatory braking curve upon approach of a stop signal

Situation C could be considered as a special case of situation B, namely one in which the speed restriction is 0 km/h, but it is more useful to regard it as a separate situation. The first reason to do so is that a correct reaction to situation C requires other behaviour from the driver than if he were in situation B, as will be shown later in this chapter. The second reason is that different accident scenarios apply to each situation. In situations A and B, a route is assigned to the train, which means the safety conditions (see section 2.3.7) are met. Derailments may occur if the dynamic forces become bigger than can be handled by the vehicle, by the wheel-rail interface or by the infrastructure. This may happen in the case of excessive overspeeding or in the case of infrastructure defects. Although some incidents of these types did happen in the last few years (Onderzoeksraad voor de Veiligheid 2005b; Inspectie Verkeer en Waterstaat 2007e), the majority of the incidents occur in situation C and are SPADs, which this chapter will therefore focus on. In situation C, no route is set behind the stopping point, which means that the required safety checks have not been performed. Overrunning the end of the route (boundary C) can lead to accident scenarios, as the infrastructure may be assigned to other users, such as other trains, road users on level crossings, an open bridge or a maintenance gang. In some cases, overlaps (buffer space) are provided behind the end of a route, but in cases where they are not provided, running even a little bit over the end of a route may have disastrous consequences.

In conclusion, in order to control the train correctly according to the schedule, the driver should thus ideally be provided with information on at least its timing, its speed and (the limitations of) the route ahead.

7.2.3 Signalling as a standard to provide information

Fixed main and distant signals
The restrictions in the speed-distance profile have to be translated into clear messages for the driver. These are the signal aspects. When all safety conditions are fulfilled and a route has been set, a driver can be *authorised* to access a track section (situations A and B) or, if no route is set, he must be told where to stop (situation C). This information is given by means of a *main signal* (see Figure 7.2). A main signal provides information about the restriction, but even so, the driver cannot see ahead far enough to see the main signal, and the restriction it shows, in time. Therefore, it is necessary to *warn* him ahead of a restriction, at or before the maximum braking distance. This is the function of a *distant signal*. It gives a preview of what the main signal is showing at that *same* moment. If the main signal shows red (stop), as in Figure 7.2, the distant signal will display the yellow aspect (prepare to stop). If the main signal allows access with restricted speed, the distant signal will inform the driver about that speed. If the main signal shows green, the distant signal will also show green, indicating that no particular action is required. Generally speaking, braking must start at the distant signal and can and must be completed before the main signal is reached. Because some time elapses between passing the distant signal and approaching its corresponding main signal, a route might have been set and the main signal can show a better aspect than what was announced in the distant signal. In exceptional cases, for example if a route has been revoked, the main aspect could also be worse than announced, which means that imminent danger is ahead.

Functions of distant signals
Distant signals are vital for the safe operation of railways, as they provide safety-critical information at a point at which the driver still has time to respond. As they are not very common in most other modes of transport and therefore may not be completely understood, we will discuss them in more detail here. The most important function of a distant signal is to indicate which restriction a driver is approaching. Basically, a distant signal says to the driver

'at some point further down the route, you will meet a certain restriction'. The questions it should be able to answer are:
- What exactly is the restriction?
- At what exact location will the restriction start?
- What should the driver do in order to meet the restriction?

If we look at the third question, it could be answered with information on how hard the driver should apply the brakes. This could be a standardised deceleration rate, but currently, no such standard is in force. The signalling system is designed such that a service braking[8] will be enough, although it is unknown exactly why this is the case. Reasons could be higher comfort, higher controllability for the driver or the fact that emergency braking rates cannot always be achieved, for instance when tracks are slippery.

Since it is quite difficult to predict the exact stopping location when a driver brakes from full speed to standstill in one go, he will (and is legally obligated to) first slow down to an intermediate speed (Figure 7.2) from which the stopping location *can* be controlled. The train should not stop too late, but neither should it stop too early, for instance with its tail on points that have to be used by other trains and/or on a location so far before the main signal that the driver cannot see whether he may proceed. Such an intermediate speed will not be necessary if the distant and main signals indicate a speed restriction, as the train can approach the signal with the indicated target speed. The distant signal not only announces a speed, but it also steers the driver's attention towards those information sources that are considered useful in the indicated scenario, as will be shown in section 7.4.2.

Classification of distant signalling information
Information potentially provided by distant signals can be classified by the richness of the information they give about the scenario: *what* should be done and/or *where* (or *when*) it should be done. An overview is given in Table 7.1.

You may think that a driver only has to be informed about the speed he is allowed to drive. This is correct in so far as this is the only degree of freedom a driver can control. However,

Table 7.1: Types of distant signalling, sorted by richness of information.

Type of distant signalling	Indication of target speed	Indication of target location	Example
Warning only	No	Not precise	Flashing orange light on motorway announcing a traffic light ahead (no rail example known)
Incomplete	Not precise	Not precise	Yellow aspect; ATB-EG cab signalling
Location specific	No	Yes (standardised in absolute distance)	Countdown markers
Speed specific	Yes	Not precise	Yellow aspect + Number for speed
Implicit	Yes	Yes (standardised distance relative to braking distance)	Fixed speed reduction sign
Complete	Yes	Yes	Cab signalling of systems such as ATB-NG, LZB or ETCS

[8] A service braking is a braking with a deceleration rate that is still considered to be comfortable and thus can be used in normal operation. This is approximately 0,6 m/s².

what is easily forgotten is that a speed limit is implicitly linked to a certain location (as can be seen in Figure 7.2) and that the driver must therefore also know *where* that speed should be achieved and where the train is currently located. This information is also required in order for him to find the next signal (route knowledge, see section 7.4).

Functions of main signals
The signalling system is designed in such a way that main signals are normally at red unless a route is set for a train. It will not normally be possible for a main signal to show a more restrictive aspect than was announced by the distant signal. One may wonder what the added value of a main signal is, if its information has already been given by its preceding distant signal. There are a number of arguments:

- Marking the point where the speed restriction starts, so that the driver can see where the restriction of the distant signal should be met. In the case of a main signal showing 'stop', the driver can derive from the position of the main signal where he must apply the brakes so as to stop short of the signal.
- Specifying details: in some cases, distant signals do not provide all required information; which can then be derived from the aspect of the main signal.
- Redundancy: in the case where a driver has forgotten the distant signal aspect, or when he did not see the distant signal (for instance when the train changed direction or when there was a crew change) the driver can derive the information from the main signal.
- Updating: it may happen that a route is set after the driver passed the distant signal or possibly even after he came to a stop. An updated main signal aspect tells the driver that he can then safely proceed.

Requisite variety of signal aspects
In an ideal situation, there is a one-to-one mapping of scenarios to signal aspects. The reason for this is simple: if there were more signal aspects than possible situations, the signalling system would be unnecessarily complicated. And if there were fewer signal aspects than possible situations, it would not be clear what the driver could or must expect ahead. In that case, the signalling system would not fulfil Ashby's *law of requisite variety* (Ashby 1956), which states that the number of regulatory actions available to the regulator of a system must at least equal the number of possible disturbances, i.e. the number of different situations that a driver can encounter. If a regulator (such as the driver) cannot be informed about all the possible situations, he will miss some or all of the information he needs to assess the situation correctly, and may therefore come up with an inappropriate control action. In that case, driving safely is only possible if the driver is able to and does choose the most restrictive one out of all remaining scenarios.

Different control behaviour at distant and main signals
The driver needs to perform DDA sequences (see section 5.3.1) at every signal. *Distant* signals normally do not require any action to be taken before the signal is passed, so that the driver can detect and diagnose the signal until he passed it as is shown on the left in Figure 7.3. Thus, it will almost always be clear whether a distant signal belongs to the track on which the train is. Additionally, in order to ensure a reliable sighting of the signal, it must be visible during a certain period of time. Most European railways use similar signal sighting times/distances; in the Netherlands this is 9 seconds at line speed with a minimum of 200 m (IRSE 1995; ProRail 2005a).

A *main* signal showing red requires an approaching driver to act differently. He will approach the signal with a low intermediate speed (max. 40 km/h on the right in Figure 7.3) and will

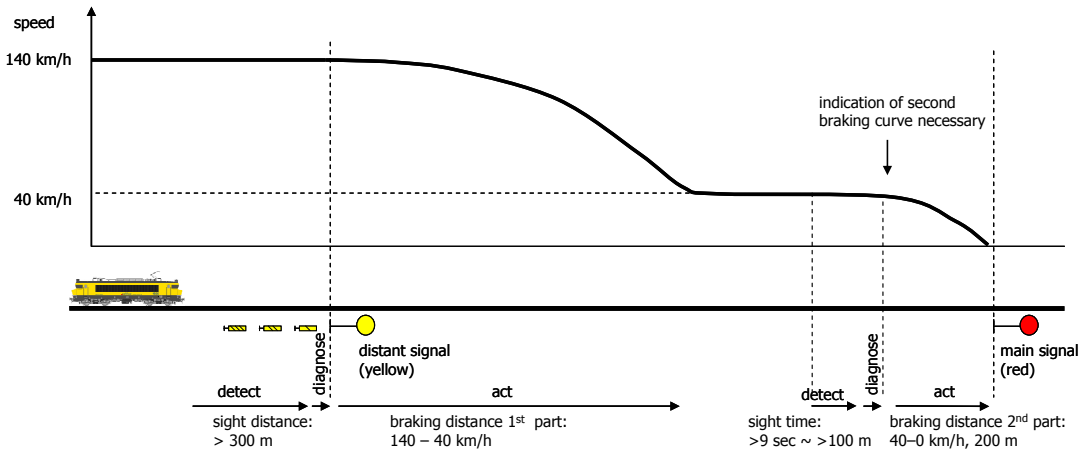


Figure 7.3: Speed-distance profile with different DDA sequences for distant and main signals.

perform a second DDA sequence in order to brake to a full stop before the signal is reached. What is different from his behaviour at a distant signal, is that in this case the action (braking) *also* has to be completed before the signal is reached, which means that detection and diagnosis will have to be done far enough before the signal, so that the driver still has time to act accordingly. We conclude that:

- Main signals and distant signals must have different requirements for their visibility distances, depending on line speed and approach speed;
- it must be clear from that distance to which track the main signal applies.

One would expect different sighting requirements for main and distant signals. Indeed, in Germany, where separate distant signals and main signals are widely used, a distant signal needs to be seen only 250 m in advance, whereas the main signal needs to be visible up to 500 m in advance (IRSE 1995). Repeater signals are required to provide a (second or third) preview if the main signal cannot be seen from 400 m distance (IRSE 1995). The Dutch design rules do not make such a difference; here, each signal must be visible 9 seconds in advance with a minimum of 200 m (AV133, ProRail 2005a).

Because main signals must be related to the correct track from a certain distance, several countries (amongst which Great Britain, Switzerland and Japan) assist the drivers by providing static or dynamic marker boards relating a signal to a track by its track number. Such information is not provided in the Dutch situation. Other ways to help the driver locate the main signal are standardised locations, auxiliary signalling and route knowledge. The first two of these, as part of the signalling system, are the topic of the next two subsections. Route knowledge is addressed in section 7.4.

Standardised signalling distances

The distance between distant and main signals can be standardised. In Germany, a standard distance of 1000 m is used on nearly all main lines (IRSE 1995). This distance may be increased to a maximum of 1500 m if sighting requirements make it impossible to place them 1000 m apart. Great Britain uses a maximum of 150% of the braking distance (IRSE 1995). This gives the driver some support in finding the main signal and leads to uniform deceleration rates. Longer distances may tempt a driver to brake not so hard, in order to prevent running at low speeds for long distances and thus being delayed. The use of standardised and maximum distances also means that not too much time passes after the distant signal has been passed before the main signal is reached. In cases where more time passes, the probability increases that the route could be set in the meantime and the corresponding main signal be cleared. Thus, long distances between distant and main signal

may lead to an increased expectation that a signal clearance will be the case, and/or the driver will have to keep the distant signal's warning in memory for a longer period of time, and may forget it. The latter is confirmed by drivers who run regularly over a track section which temporarily has a maximum speed of 40 km/h instead of 80 km/h, so that the distance between distant signal and the stopping location is too long. This requires drivers to brake only a long time after passing the signal and remember that during this time.

Auxiliary signalling
In the case of lineside signalling, auxiliary signalling is provided to help the detection of (distant or main) signals, often in the form of reflective countdown markers. Auxiliary signalling therefore falls into the category of location-specific information. Note that the countdown markers at distant signals help to detect the signal, something they do not do in the case of main signals, as they are placed at a point where diagnosis should already have been performed (see Figure 7.3). Nevertheless, they do provide an indication of the remaining distance to the signal, which could be useful, but countdown markers are not used in most stations.

From DDA theory it might be expected that the driver's task would be eased if the point where he must start to brake from 40 km/h to standstill would also be indicated with a second distant signal (the location at which the action of the second DDA loop has to be started in Figure 7.3). This would have the following advantages:
- Standardised location relative to the main signal;
- Updated preview of the main signal (if this is not visible);
- There is no longer a need to locate the main signal *ahead* in order to stop in front of it, as both its aspect and its remaining distance can be derived from the intermediate signal;
- Red signals can then be approached in almost the same way as other signals.

Although no generic application in this way is known, such a signal is now being placed near Amsterdam Muiderpoort. This solution was chosen as signal number 440 was repeatedly passed at danger: a combination of signal positioning in a curve and the view being obstructed by overhead line equipment meant that drivers could not establish a reliable connection between the track they were running on and its signal (Inspectie Verkeer en Waterstaat 2007b). Furthermore, the configuration of the automatic route setting system was such that access to this track was always given with a yellow aspect in the previous signal, thus warning for a stop at this signal 440; but in practice, this signal had nearly always been cleared once the driver approached it, so that drivers were likely to learn that 'prepare to stop' at this location in practice meant 'do not prepare to stop'. The new signal at the braking distance before the main signal will allow access to this track with a green aspect at all times and provide up to date information on whether or not signal 440 has been cleared at a point further down the route; also it removes the need to locate main signal 440 ahead. The additional distant signal thus takes away both the update and the visibility problems. The main signal's function then is only to reflect updates once the train has come to a full stop. Such (repeater) signals are commonly used in Germany (see earlier).

The main point of this section was to indicate that two functions of railway signalling must be clearly distinguished: distant signalling, which provides information about where to ultimately brake (i.e. controllability envelope) and main signalling, which indicates the limits of the movement authority (in terms of a speed-distance-profile, i.e. an operational envelope). This means that a driver has to perform tasks differently depending on which of the two functions he must fulfil; this also poses different requirements to the two types of information being presented.

7.3 The Dutch signalling system and its practical use

7.3.1 Design principles

Speed signalling system
This section will show how the foregoing signalling principles are embodied in the Dutch signalling system. The Dutch signalling system was developed in the 1950s. At the time, a large number of existing signalling systems were analysed and the best elements selected. An important conclusion was that it should be a speed signalling system, contrary to many former signalling systems which indicated the route rather than the permissible speed. In a route signalling system, a distant signal would indicate for example 'you will be diverted after passing the next signal', leaving the appropriate speed up to the driver. In a speed signalling system, all the driver has to do is compare his current speed to the indicated speed. Many drivers indeed indicated that they find this system easy to use. However, as only 'speed' and 'route' appeared to be considered during the system's design, it seems that variables such as 'distance' did not receive that much attention.

Requirements for the signalling system
Amongst others, the following requirements were formulated for the system (AV130, ProRail 2005a):
a. The signalling system must be able to indicate all circumstances that are relevant for the driver in order to regulate the speed of his train.
b. The signalling system must be able to indicate at what speed the train is allowed to drive.
c. A prohibition to proceed must be announced at least at the braking distance in advance of the signal that shows stop.
d. An order to decrease speed must be announced in such a way that the decreased speed can be achieved at the right location and by means of a service brake deceleration.
In addition, it is concluded that '*each indication requires a separate signal aspect.*' These requirements are in line with Ashby's Law and provide the requirements necessary to be able to react in time, as can be derived from the DDA model. However, even though requirement (d) states that a decreased speed must be achieved at the right location, the signalling system is not specific about *where* that location is. This is what we defined as implicit distant signalling. Section 7.3.3 will show that requirement (a) is not met in all circumstances.

The Dutch rail traffic regulations (part of the railway law) also put forward the following requirements, amongst others (*Regeling Spoorverkeer* 2004):
article 26: Signals shall be placed, or marked such that it is clear for the driver which signals are meant for the track over which he will run.
article 27: Signals shall be placed such that the driver, depending on the maximum speed, is able to perceive them in time and to react to them appropriately.

These requirements are well in line with the requirements to perform a DDA sequence.

Unified main and distant signals
Prior to the development of the signalling system, a block length (a block is defined as a length of track between two main signals that can be assigned to one train) of about 1500 m became more or less the standard. As this length is somewhat similar to the required braking distance, the designers of the signalling system decided that each signal would become the distant signal of the subsequent one, and that there would be no more separate distant signals (de Vos tot Nederveen Cappel et al. 1964). Each signal shows red if the block behind it cannot be accessed, yellow if the block behind the signal is clear but the subsequent one is not, and green if both this block and the subsequent block are free. This is summarised in Table 7.2.

Table 7.2: Meaning of Dutch signal aspects compared to separate main/distant signalling.

Free blocks	Dutch Aspect	Required action	Separate main/distant signals	
			Main aspect	Distant aspect
2	Green	no action necessary	Green	Green
1	Yellow	brake to 40 km/h, prepare to stop	Green	Yellow
0	Red	stop	Red	(n/a)

The integrated signal incorporates both functions of main and distant signals. If it is showing red, it clearly acts as a main signal and no distant information is given. This information is irrelevant anyway if the signal must not be passed. If it shows green, it acts both as a main and a distant signal, indicating that no restrictions apply at this signal nor at the next one. If it shows yellow, it has the same meaning as a green main signal (access to the block is given) and a yellow distant signal (access to the subsequent block is not given). Because the yellow aspect is considered to be more important, only the yellow aspect is shown. This necessarily means that signals showing a yellow aspect (with or without a number) never give information about the speed at which they themselves may be passed. Both the green and yellow aspects can be supplemented with a number giving a speed indication. A number shown with a yellow aspect indicates that the train must decelerate to the speed indicated by that number (e.g. yellow+8 means slow down to 80 km/h). If the train must come to a full stop (i.e. 0 km/h) no number is indicated, so a number with a yellow aspect makes the signal aspect less restrictive. If shown with green, the number indicates that the allowed passing speed at the signal is lower than the maximum line speed. Therefore, the green aspect is flashing if shown with a number.

Eroding knowledge of distant signalling
Despite the rationale to integrate main and distant signals, separate distant signals still exist. Now that distant signals *as objects* are not so widely used, it appears that knowledge about their *function* is also gradually eroding. In interviews, some professionals who advise on related issues, could not tell the difference between a main and a distant signal. The Dutch legal definition is remarkable as well, because it only defines a distant signal by what it cannot do (*Regeling Spoorverkeer* 2004):
- *main signal: light signal that can emit red light;*
- *distant signal: light signal that precedes a main signal and cannot emit red light.*

This of course is true, but it is not very useful to define the function of an element by what it *cannot* do. Of course, the essential characteristic of a distant signal is that it (only) provides a preview of the main signal ahead. Drivers are supposed to react to distant information rather than to a main signal, so it is evident that this information should be clear. This information forms the basis of safe railway operation and erosion of distant signalling knowledge is worrying.

In most public SPAD incident analyses by the railway inspectorate IVW, the focus is on the (main) signal that was SPADed and not on the preceding (distant) signal(s) at braking distance, so that possibly confusing situations at the distant signal may be missed in the analysis. Similarly, the current SPAD reduction program by the Dutch rail sector, signal visibility receives relatively much attention (Verheij and van Dijk 2008), but this approach appears to skip the fact that drivers have to be informed primarily by distant signalling and not (only) by seeing the main signal.

As will by now be clear, the two different functions of distant and main signals are now physically combined in one device. Although the functions remained the same, this

integration leads to difficulties in some situations, as the remainder of this section will show. These difficulties have their roots in the fact that the distance between two successive signals is highly variable, especially in station areas, and in reduced message variety.

7.3.2 Determining the most relevant speed profile

Although visibility is important, an even broader approach is required from a driver's viewpoint. To perform a DDA loop, not only signal visibility is important, but also the 'understandability' of signals in order to diagnose the situation and find out which of the perceived information requires the most restrictive behaviour. Thus, the driver must be able to determine the most relevant speed profile. A driver approaching the station in Figure 7.4 on the left track, has to define the most relevant speed profile from the following bits of information, each indicating (part of) a speed profile.

- deceleration from line speed to 80 km/h, to be finished at the square '8'sign;
- deceleration from 80 km/h to 40 km/h, starting at the triangular '4'sign;
- deceleration from line speed to stop, to be finished at the red signal that is partly obscured by overhead line equipment, the station building and vegetation beyond the platform end;
- deceleration from line speed to stop, to be finished at the end of the platform (for local trains only);
- maximum speed for rolling stock type.

Even when the driver has determined the most relevant speed profile, he must still keep the limitations that are not critical in memory. For example, if a station stop was most critical, the driver must remember that he was also approaching a signal showing a stop aspect. Confusing several braking plans may lead to safety risks (Hill et al. 2007).

The examples that will follow in this chapter show that where things go wrong there is either information missing, from which to derive the most relevant speed profile, or that there is an information selection problem. In the latter case, there is too much information about all kind of speed profiles, so that the driver might not discern the information related to the most relevant speed profile and irrelevant information is accidentally identified as most relevant. As British research revealed, the physical co-location of information per se does not constitute problem as long as the cognitive processes of the driver in order to process the information are not co-located in time (Hill et al. 2007).



Figure 7.4: Several speed profiles; Zaandam Kogerveld.

7.3.3 Difficulties with the yellow aspect

No passing speed indication in yellow aspects
Although the Dutch signalling system was designed as a speed signalling system, it does not give any speed information for speeds lower than 40 km/h, even though this may be useful to construct the most relevant speed profile and thus to regulate the train's speed. Also, all yellow aspects (possibly accompanied by a number) do not indicate a passing speed. The driver can normally derive this speed from the signal aspect in the previous signal, but there are some reasons why this information may be necessary in an explicit way:

- The yellow aspect does not give information about the (variable) braking distance to the next signal, hence no passing speed can be derived from the signal itself.
- The previous signal may have been not passed by this driver, e.g. when the driver has just taken over his shift or when the train reversed its direction.
- The previous signal may have indicated information that is no longer up to date, for instance when a route has been set in the meantime.
- The driver forgot the previous signal aspect, e.g. due to a station halt.

Although the Dutch train protection system ATB-EG alleviates these drawbacks in many cases, these circumstances also manifest themselves in station areas, where ATB-EG also provides no support for speeds below 40 km/h. In all cases where this information is missing, a signal showing a yellow aspect can be followed by a red signal within a very short distance, and a driver should be able to stop on sight. However, he is not comprehensively instructed to do so; neither by the signal aspect, nor during the current driver education.

Twisted logic: yellow – yellow – red aspect sequence
The logic of the unified speed signalling system is very clear on the open line (running line between two stations). In station areas, it may be difficult to predict where the next, red signal will be, because in stations situations may occur that are unusual or impossible on the open line. To illustrate this, Figure 7.5 presents a typical station layout. Firstly, the signal spacing in a station is far less regular than one every 1500 m, because signals are not only required to space trains but also to protect points and to enhance shunting manoeuvres, so their location is dictated by the track layout. Secondly, the driver can be directed to a number of possible routes. This will be examined in more depth in the next section.

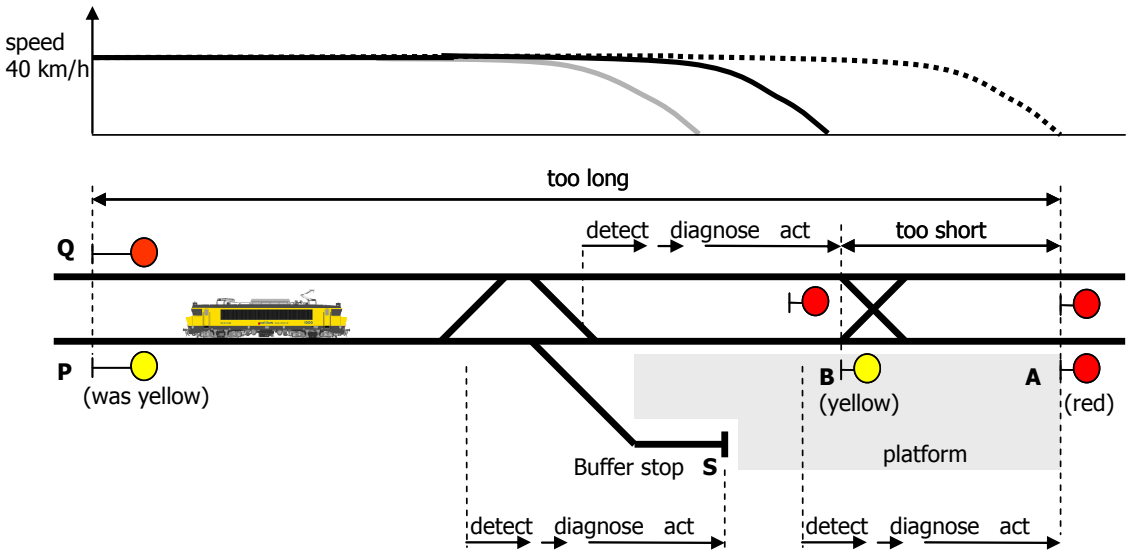


Figure 7.5: Typical situation of an ambiguous yellow aspect.

In quite a few cases, signals are spaced closer than the minimum required braking distance of 400 m (at the lowest speed of 40 km/h that can be signalled). Examples can be found at nearly all major stations, where points divide platform tracks into two parts so that several trains can be handled at the same platform. In those cases so-called yellow-yellow-red aspect sequences occur, which provide ambiguous information to the driver. By the logic that the last signal before a red one should show yellow, signal B (at a too short distance before signal A in Figure 7.5) shows yellow. However, according to the logic that a driver should be warned at least at braking distance before the red signal (a requirement which signal B does not meet), signal P will also show yellow. In these cases, both yellow signals are ambiguous. Firstly, signal P does not tell the driver whether he should stop at the next signal (B) or the second-next one (A). In former times, signal P would therefore show a flashing yellow aspect to warn the driver if he *normally* had to stop at the second half of the platform, and against his expectations would now get a route that is more than 200 m shorter leading only until the first half of the platform. Secondly, the second yellow aspect (B) does not provide clarity on the available braking distance between the yellow and the red signal, so that it remains unclear what a safe passing speed of that signal is.

If signal A is positioned within braking distance of signal B, the most relevant speed profile is determined by signal A rather than by signal B. This situation, where the driver is required to look ahead *further* than the next signal, is unusual because the design of the signalling system was based on the assumption that each signal would be the distant signal of the next one. In this case however, it is not. Nor is the driver told explicitly that he may have to look *through* the next signal. This type of distant signalling is incomplete, so that further information will have to be derived from the stop signal itself. It leaves the driver with a problem to correctly diagnose the situation. It is therefore extremely important that the red signal itself in such cases is visible from at least braking distance.

Different routes indicated by the same yellow aspect
The signalling system neither indicates which route is set, nor where it ends. This situation is always indicated by a yellow aspect, regardless of whether a route is set to signal A, signal B or buffer stop S in Figure 7.5. This may lead to problems if the braking distance is significantly shorter on one route than on another, e.g. when the train is led to the buffer stop and not to signal A, especially if a driver is not normally led to the shorter track. It requires the driver to consider all possible routes and identify the most restrictive one. This task however is much more difficult than normal behaviour on the open line, as it requires the driver to know all possible routes and the location of the corresponding signals and then to draw conclusions about a safe speed. This is even more difficult in combination with yellow-yellow-red aspect sequences because in some cases a driver has to react to a red signal before passing its preceding signal with a yellow aspect.

The additional information the driver needs in this case could be supplied by the traffic controller, who knows which route he will set or has set. However, it was discovered from investigations that SPADs occurred because drivers were distracted when they were called during driving. It was decided to prohibit calling when the driver is driving without backup of the ATB system, which is when he is driving slower than 40 km/h and/or if a yellow ATB cab signal is displayed (NS Reizigers 2006). Although this measure could be effective to prevent the driver from irrelevant communication (such as a roster planner asking about something that is only relevant in 3 hours), it also cuts off the information that could help the driver to perform his task in the situations described above. It is somewhat paradoxical that a driver is not allowed to be informed about a danger *because* he does not have a safety net; for exactly the same reason, it could also be argued that it is exactly those routes that he should be

informed about. Interviews revealed that traffic controllers in the formerly integrated NS organisation were obliged to inform the driver if they set a route which was shorter than the driver might expect, but this is no longer so.

Again, it is essential that a driver can see all possible endpoints of his route. However, as the required visibility distance is only 200 m, situations exist in which there is no time to detect and diagnose the signal for trains having a braking distance of up to 250 m (from 40 km/h). It is questionable whether such layouts comply with article 27 of the rail traffic regulations (*Regeling Spoorverkeer* 2004).

Possible meanings of the yellow aspect
We can conclude that unifying distant and main signals can lead to comprehension problems: the design choice to integrate them led to the position of distant signals being (implicitly) dependent on the position of the foregoing main signal, although it has no logical relation with it. The yellow aspect is used in essentially different situations, in which it can have the following meanings:
- Intended signal sequence: the next signal is red and at least at braking distance (in Figure 7.5 signal P announcing signal B or buffer stop S);
- Short block announcement: the next signal is not red, but after the signal a red signal will follow within braking distance (so the driver should look *over* the next signal; signal S announcing signal A);
- Short block: the next signal is red and within braking distance (so the driver should be braking when passing the signal; signal B announcing signal A).
- In each of these cases, the yellow aspect may also announce a yellow flashing aspect instead of a red aspect.

This analysis demonstrates that the signalling system does not show the requisite variety to clearly differentiate the various scenarios to the driver. We can conclude that the yellow aspect neither signals a speed, nor a route, nor a distance to the next signal, in fact requiring the driver to drive on sight. This was indeed the formal meaning when the signalling system was introduced, but it has been abandoned during the past decades as will be shown in section 9.4.4.

The Dutch signalling principles have apparently been copied by the Swiss railways (Bischof 1986), although they have been slightly adapted to cope with these anomalies in stations. The Swiss have added additional indications to both yellow signals (*Schweizerische Fahrdienstvorschriften FDV (R 300.1–.15)* 2006), adding information about the actual situation and thus complying with the law of requisite variety. Although a similar proposal was made in 1987 (Middelraad 2000), such changes were never made in the Netherlands. It is not surprising that this was a factor in quite a few accidents, for example in the past few years in Amsterdam (2004), Rotterdam (2005, 2006) and Arnhem (2006). What may be even worse, is that none of the public accident reports considered this shortcoming as one of the causes of the accident, probably because the signalling system is considered as given (see Van den Top and Groeneweg 2007 for an analysis of the Arnhem accident). It is to be expected that more accidents of this kind will happen as long as the ambiguity in the yellow aspect remains.

7.3.4 Difficult situations with signal detection from a distance

Ambiguity in signal information, as explained in the previous sections, leads to difficulties in determining the most relevant speed profile. However, it is the task of the signalling system to inform the driver about that profile. The yellow aspect is the last warning before a full stop and can be considered as the most important signal aspect (de Vos tot Nederveen Cappel et al.



Figure 7.6: Example of a situation where a single signal could be related to two different tracks.

1964). However, it is also the one that provides the least clarity about what the driver can expect (see Table 7.1 and Figure 7.5). It is therefore essential that the driver is educated about the difficulties that he may encounter, that he can notice the red signal in time, that the view to the signal should not be obstructed, and that the driver can make a clear connection between the signal and the track he is running on. Figure 7.6 gives a striking example of how difficult it can be to relate a signal to the correct track even if the view is clear. The photograph (identical left and right) is taken from the location where the driver should start braking. The dotted lines, which the driver draws in his mind to relate the signal to a track, indicate that the same signal could be positioned next to either one of two possible tracks, if it were placed on a pole.

In interviews with drivers and in analyses of SPAD incident reports, the following situations have been found in which detection of the correct signal (that is: the one applying to the route for the train) can be difficult, either because it is difficult to detect the correct signal or because it is difficult to diagnose the current situation:

- *Signals on gantries*. A reference to the track is missing, especially in curves and when there are many tracks (see Figure 7.6).
- *Dwarf signals*. Dwarf signals are in some cases meant to enhance shunting manoeuvres and are then placed at places where train movements would not normally be brought to a halt. Visibility problems include that a train on the neighbouring track can obscure the signal from a clear view in right curves. Some dwarf signals are placed on short posts (meant to increase visibility) and are then sometimes wrongly perceived as being further away than they actually are, because they are seen closer to the horizon.
- *In curves*. Sometimes signals can be seen only from a short distance, or from a longer distance it can be unclear to which track they apply (see Figure 7.6).
- *Diverging points shortly before the signal*. It is not clear onto which track the train will go, which means it is also not clear which of the signals is relevant.
- *Situations in which there are signals for some, but not all tracks*. This will go wrong especially when there are many tracks so that drivers have to count signals and assume that, if they are on the third track from the left, the third signal from the left applies. This will go wrong if some of the tracks do not have a signal.
- *Signals which have a short sighting distance*. Some signals are not clearly visible from the point where a driver should start to brake, e.g. when they are placed in a curve or when overhead line equipment restricts their view.

Figure 7.7 shows an example of a situation where all these factors play a role at once.

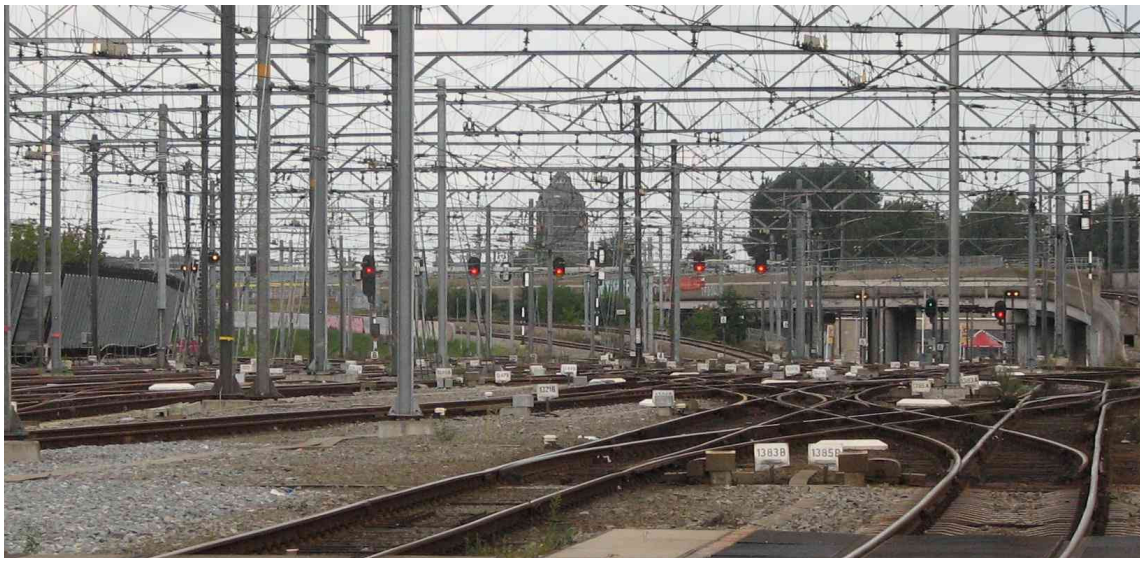


Figure 7.7: Example in which many difficulties for drivers coincide; Utrecht Centraal.

7.3.5 Difficulties in determining the most relevant speed profile

From interviews drivers said they have little problems in the following signal approach situations:

- with *normal green-yellow-red signal aspect sequences*;
- when running at *green* signals, drivers know that there is no need to perform any action until new information indicates so;
- with signals *changing from red to yellow because another train is moving in front* of their train.
- With *planned signal stops*, e.g. at stations.

However, the following situations have been identified in which determining the most relevant speed profile can be difficult:

- The *yellow-yellow-red signal aspect sequence*, which was analysed in section 7.3.3; especially in stations, drivers may be looking for train length indicators so that they can stop their train at the correct position along the platform. This may lead to a SPAD if there is an unexpected red signal halfway along the platform. This is often part of a yellow-yellow-red sequence and therefore not clearly announced in the distant signal.
- *Departure on a yellow aspect* (see also section 4.6). This means that a driver departing a station receives his last warning for a stop signal at the moment he is still accelerating, but the yellow aspect gives little information to hold on to when it comes to which speed should be accelerated to or where to stop. It is contradictory to accelerate with a signal aspect that says that a driver should prepare to stop, and no information is given at the moment that the driver should start braking. Since drivers are generally always warned at the point where they have to brake, this is exceptional.
- *If a yellow aspect, indicating an unplanned stop, is given at a point where a driver has to slow down also in the normal scenario.* The driver may be likely to commit a routine error and slow down to the speed that is appropriate in the normal situation, instead of braking harder and to a full stop. For example, the SPAD workgroup performed detailed analyses of repetitive SPAD incidents near the small towns of Weesp and Boxtel, which seemed to be attributable to this mechanism (Carstens et al. 2007). In both cases, the driver must normally slow down to 80 km/h, but incidentally to a full stop. This requires much harder braking, which was in some cases not actually done and SPADs resulted. A driver could not easily detect his error, since in both cases the stop signal was placed after a curve. Also

the ATB system, clearly meant to mitigate driver error, will not intervene in this case. After lengthy debate, it was decided that it should only check whether the driver is braking but not whether that is hard enough (Studiecommissie ATB 26 oktober 1967). Combining speed reduction signs and signals, as was the case in Boxtel, saves investments in the ATB equipment, but makes human error more likely. It is a somewhat paradoxical coincidence that the human error probability is inadvertently increased in order to save expenditures on measures to mitigate human error. In order to understand SPADs, we learn that it is important to look not only at the course of events that led to the SPAD, but also to compare that to the course of events in normal situations. If these are similar, routine errors are likely. In other words, the warning for a stop signal should under all circumstances be very explicit and different from other instructions.

- *If static speed profiles (indicated by fixed signs) and dynamic speed profiles (indicated by signals) overlap*. One driver explained: *"sometimes there is a fixed speed reduction to 40 km/h [for example before a curve]. When you are slowing down to 40 km/h, you encounter a yellow signal. Now, you learn in your education that signal speeds are more important than speeds indicated by signs. So your main attention is then to stop at the red signal, but then you can forget that you must have reduced your speed to 40 km/h at an earlier point along the route"*. This is an example where a driver's attention is focussed based on the information perceived, and inadvertently two speed profiles are confused. This will be elaborated on in section 7.4.2 about driver expectancy. Also, it makes clear that education, driver expectancy and infrastructure design cannot be considered in isolation.

It would be interesting to find out how the SPAD probability is influenced in these cases, since it might give a clue on driver performance if one could manage to make unplanned stops also less unexpected, for instance by providing information about the necessity for an unplanned stop before the driver reaches the yellow signal. In that case, the signalling system would again fulfil its intended backup function and would support the driver's expectation.

7.3.6 Driver's indications confirmed by data

Confirming driver's perceptions

The previous two sections provided a list of situation descriptions in which drivers said they have problems with signal layouts (section 7.3.4) or with signal approach situations (section 7.3.5). These situations are summarised in Table 7.3 and Table 7.4. It is interesting to check whether there is evidence or counter-evidence for the accuracy of driver's perceptions, coming from other sources, such as SPADs. To do so, it is necessary to calculate the SPAD probability per approach (# SPADs divided by total number of approaches) when the signal is showing a red signal aspect, and to do so for each of the situations described before.

A normal signal type (i.e. signal on post alongside a straight track) and a normal signal approach situation (i.e. a green – yellow – red aspect sequence) are used as a reference factor. The aim of the analysis in this section is to find multiplication factors for different situations, in order to be able to make statements in the form 'the SPAD probability when approaching a dwarf signal is 4 times higher than the probability when approaching a normal signal'.

To start right away with the answer: calculating such factors is only possible to a very limited extent (for 4 of the list of signal layout types in Table 7.3). This is not because the calculation is difficult, but because there are severe limits to data availability. What is available, however, does not contradict the driver's perceptions.

Table 7.3: Multiplication factors for various signal layout types.

Properties of signal showing red aspect (see section 7.3.4)	SPAD probability multiplication factor
Standard signal on post (= reference)	1.0
Signal on gantry	approx. 2.0
Dwarf signal	approx. 4.0
Signal in or behind curve	approx. 1.7
Signal <250 m behind points	approx. 1.5
More tracks than signals	?
Short sighting distance	?

Table 7.4: Multiplication factors for various signal approach situations.

Approach situation of signal showing red aspect (see section 7.3.5)	SPAD probability multiplication factor
Normal green – yellow – red sequence (= reference)	1,0
Previous two signals cleared to yellow on approach (train moving ahead)	?
Red signal expected (e.g. planned stop)	?
2nd yellow in a yellow-yellow-red sequence (unexpected stop)	?
Red in a yellow-yellow-red sequence (short signal spacing)	?
Yellow distant signal aspect coincides with a usual speed reduction (habituation)	?
Departure on yellow aspect (proceed and stop message given in one aspect)	?

Table 7.4 sums up a number of signal approach situations that drivers perceive as difficult. However, for none of these situations were data available to calculate the SPAD probability multiplication factor. Apparently, the railways try to understand SPADs from static properties such as signal properties, and not from the dynamic situation in which the driver plays a role. In fact, during this research, 'unplanned stop' was added to the inspectorate's SPAD data collection form, so that some of the missing information will now be collected.

To calculate the multiplication factors, a database was used from the sector-wide SPAD workgroup (section 2.2.7). This database was set up by the railways according to the method described by the SPAD workgroup (Wright and Vorderegger 2005). Appendix D contains a detailed description of the calculations made. In brief, this appendix indicates that:

- the database does not contain information on how often signals are approached with a red aspect;
- the database does not contain data for all criteria mentioned by drivers, such as sighting distances or signal approach situations;
- not all entries in the database contained enough data to calculate the probabilities;
- a definition problem for some fields in the database made it difficult to interpret the data, and these fields have been left out from the calculation.

This means that the results should be interpreted with care. Further cautions in interpreting the data are:

- The calculated error probability may still have a considerable variance (that may or may not be explained by some of the factors marked with a question mark in Table 7.3 or Table 7.4), so on certain spots the SPAD probability may be far higher or lower.
- The data should mainly be considered as a relative measure, e.g. to approximate that dwarf signals on average have a 4 times higher probability of being SPADed than a signal on a normal post.

Results

Table 7.5 shows the resulting SPAD probabilities for each signal type and layout. Note that these results include also approaches at other than red aspects, so that the real SPAD

Table 7.5: SPAD probabilities ($\times 10^{-6}$) per approach at either one signal aspect for different signal types and layouts.

	no. of signals	average	in curve ?		<250 m behind points?	
			no	yes	no	yes
signal on post	1383	1.8	1.7	2.7	1.8	3.0
gantry signal	309	3.4	2.8	5.6	3.1	8.9
dwarf signal	759	7.9	7.4	12.0	6.9	9.9
total	2451	2.9	2.7	4.3	2.5	7.3

probability must still be divided by the share of red aspects shown in the signal. For example, if two in a million approaches result in a SPAD, and the signal shows red for 10% of the approaches, the SPAD probability per approach at red is 2 in 100,000.

Despite some restraints and uncertainties, we can still use the data to see whether there is reason to doubt the drivers' perceptions. The following conclusions can be drawn.

1. This information gives no reason to doubt the validity of the drivers' perceptions.
2. The sector-wide SPAD workgroup used a database that has considerable flaws and as such only gives limited possibilities to find out which factors increase the SPAD probability. Instead, the database was used to install technical measures (ATB-VV) at the most SPADed and most risky signals. Although this may increase safety, it does not help to understand why the SPADs actually occur, thus leaving the difficult situations for drivers in place. It now also appears that the ATB-VV measures will not have an effect as large as expected (Inspectie Verkeer en Waterstaat 2008h), so that the question rises whether the equipment has been installed at the correct locations.
3. This database or a similar one could easily be used to calculate SPAD probabilities per approach, but that requires that signal aspect data are added. If a distinction is to be made between different signal approach situations, the situations as mentioned in Table 7.4 should also be added.

Also in Chapters 8 and 9 we will find explanations for the problems mentioned by drivers.

7.4 Driver human factors

7.4.1 Introduction

Now that we have seen which situations are difficult for drivers and how this relates to the signalling system's design, we will try to understand why exactly these are difficult situations. To do so, we will look into human factors. This section will start with driver expectancies and driver distraction, which are generally considered by the railways as frequent causes for SPADs with 19% and 17% respectively (Inspectie Verkeer en Waterstaat 2007f). Then, we will look into vigilance and route knowledge, which are often considered by the railways as solutions for SPADs or at least as prerequisites to prevent them. We will see that none of these four terms is clearly defined or understood by the railways, which makes it difficult to formulate sound solutions. This section aims to provide a beginning of a clarification of the topics and concludes with some recommendations for better supporting the human operator.

7.4.2 Driver expectancy

Driver expectancy as a cause for SPADs

Driver expectancy is mentioned by a large number of different authors in relation with SPADs (Middelraad 2000; Griffioen and Vorderegger 2001; Onderzoeksraad voor de Veiligheid 2005a; Sanders 2005). However, not so much is known about expectancies and the word is

used in a number of different settings. In several SPAD incident reports, the word was used for different situations, such as:
- the driver expected to be sent to another track than was actually the case;
- the driver expected that a signal normally is or would be cleared where it in fact was not;
- the driver did not expect a signal at a certain location; or
- a driver expected to be allowed to depart before or after another train he could see.

Because they see it in accident investigations, the railways apparently concentrate on expectancy as a negative human property. The term expectancy is then used by the railways to indicate a difference between the actual situation (that is objectively 'true') and its representation in the driver's mind (which then is false). Having an expectancy thus appears to be considered as a negative human property that leads to accidents. That is not true, as will be shown in this section. The positive side of expectancies is hardly ever acknowledged by the railways, or if so, it has another label such as route knowledge or rolling stock knowledge – which in fact are expectancies about what the route ahead will look like or how the train will behave.

Expectancies help to perform tasks efficiently
In the world around us, there are many cues and signals. Although they are all present, only some of them are relevant to the task we are currently performing, and the non-relevant cues can be ignored. This was already mentioned briefly in section 5.3, where it was stated that an operator's expectation influences when and where he will look for new information. This subsection will elaborate somewhat further on this topic. It is not only valid for drivers, the main topic of this chapter, but also for traffic controller's, whose work will be addressed in the next chapter.

Hale and Glendon (1987) write that: *"skilled routine behaviour demands that information coming in from the outside world is processed very rapidly (…). This means that only very selective features of the situation are attended to. It is as though people are only checking that things are as they expect them to be. 'The next step of the action sequence demands that the situation should be in state X: let us check whether it is state X or not (…)'. The rest of the mental picture of the outside world is then filled in with what computer jargon calls 'default values' "*. Thus, the human brain filters information and tries to confirm what is already expected. Striking information is processed further; information that was expected is attended to cursorily or even neglected. What counts is the 'surprise value' of information. The brain does not primarily assess information in a logical way, but it mainly checks whether information matches certain patterns. That is efficient rather than stupid behaviour (Klein 2004): having an expectation enables operators to work quickly and efficiently. It helps people to take the correct action at the right moment, as also rail human factors experts indicate (Weeda 1999). Also Mica Endsley's concept of situation awareness (1995), which was introduced in section 5.4.3, consists of three parts of which the third one clearly involves producing an expectancy: *'Situational Awareness is:*
1. *the perception of elements in the environment within a volume of time and space,*
2. *the comprehension of their meaning, and*
3. *the projection of their status in the near future.'*

An operator who has to monitor a certain parameter value will look at the value at some time interval; the length of this interval is dependent on the expected rate of change of the parameter value, says Van Westrenen's alarm clock model (van Westrenen 1999). For instance, a captain navigating in a busy canal will look frequently how other ships are moving, whereas a captain navigating on an empty sea will look not so often. This shows that

expectancy, the third step of the situation awareness model, also governs the first step of Endsley's three steps, being the search for and perception of new information. An expectation serves to focus attention efficiently towards only those places where it is needed, thus freeing up mental capacity to perform other tasks. Imagine what would happen to someone who would not have the ability to have an expectancy. This would lead to fully reactive behaviour, thus creating the need to always focus attention towards everything (because everything would come unexpectedly by definition) and thus to very high stress levels – which again leads to higher error probabilities (Kirwan 1994). Although expectancies are certainly helpful to perform tasks efficiently, they also have a drawback. An expectancy, regardless of whether it is correct or not, serves to determine where further relevant information will be sought and can be found. A false expectancy can propagate itself in this way, as information that clearly tells that the situation is different than expected can be missed. Two mechanisms to explain this have been found from literature:

1. information is not perceived (even when it is looked at) when it is already believed to be known. This is the default values problem.
2. information is easily missed when it comes unexpectedly (e.g. because it is not task-related) and is not paid attention to, even when it presents itself clearly. This is called inattentional blindness.

Default values problem

Endsley (1999) found that having an expectancy with default values can also be a source of situation awareness problems, when operators *'tended to rely on habitual expectations about the behaviour of the system rather than current data, even when it was available'*. What Endsley observed during real operations, was confirmed by a simulation study with road users (Martens and Fox 2007). It was found that familiarity with the route increases rapidly. Car drivers already showed strong expectancies of a situation after having driven a route only five times, and did not notice that traffic signs were changed, even though they glanced at them. Expectancies are so strong that *'information that is highly relevant for the driving task is most likely not being perceived or at least not responded to if it does not fit the expectations'*. This was also concluded by Jongerius (1993) in his comprehensive book on Dutch railway accidents: *'(this accident) shows that deviations from normal routine are frequently not observed. What one is not used to, is not eye-catching unless one's attention is especially drawn towards it. This is something that also plays a role in many later accidents (…)'*. Default values can take over from actual information. Even if newly perceived information is actually in conflict with that picture, it may happen that information is considered to be incorrect or is distorted to fit, instead of being used as a trigger to adjust the mental picture (Hale and Glendon 1987). An example is that of a traveller who may think that a certain side road is not shown on the map, instead of using that information to conclude that he is either looking at a wrong part of the map or that the map is correct and he has lost his way.

Inattentional blindness

Other research shows that information that is not expected can easily be missed or that it takes a long time until it is noticed, even if the information presents itself prominently to the observer. A stunning example is an experiment in which people watch a video in which two baseball teams pass a ball. The observers have the task to count the number of passes in either the black or the white team. During the video, a gorilla walks clearly through the scene. Only about half of the observers notice this event (Simons and Chabris 1999). This phenomenon is known as change blindness or inattentional blindness. It shows that visual perception of change in an object or a scene occurs only when that object or the scene is given focussed attention (Rensink et al. 1997), which again requires that the observer must realise where to focus attention in order to see the change.

Understanding situational awareness to improve performance

The previous sections tried to explain how expectations are formed, to make clear that they are an inextricable part of human behaviour and that this helps operators to work efficiently, but it also showed what can go wrong with expectations and why. 'Driver expectancy' is not only a cause for SPADs, but also the reason that drivers are able to stop correctly in most cases, because they do expect a stop signal and because they do know the braking characteristics of their train. Better understanding of situation awareness will allow for better operator support. This means that it is of utmost importance that drivers (or operators in general) get clearly informed about the future situation especially if the situation is different from the expected one. This is why it is so important to know whether the infrastructure layout was changed (change in route knowledge), whether the braking performance of the train is reduced or unpredictable (change in rolling stock knowledge) or whether another route has been set (change in movement authority).

In fact, most SPAD accidents can be considered as a situation in which a driver is executing a different plan than that of the traffic controller, thus making SPADs primarily a communication problem. It is not necessarily said that the traffic controller's plan is always correct, because it sometimes contains conflicts (see Chapter 8). If the traffic controller were to communicate his plan to drivers (instead of just the result via the signalling system) the signal aspects would no longer come unexpectedly and a double check would be possible. This would also allow drivers to detect any discrepancies between the plan and the signal aspects, which would have a high surprise value and thus be given attention. If a driver's expectation is wrong, this can be classified as 'human error' but it should also be checked whether the information provided was by any chance not clear enough (and can thus be misdiagnosed) or even that necessary information was absent. Clearly, expectancies are a human property and wrong expectancies can surely never be eliminated. But they are influenced by the information provided and, as this information is often given via technical systems, there is a distinct link between expectancies and the design of the systems the operators have to use. In aviation, situation awareness has become a key element to strive for in operator training and in systems design, both for air traffic control and for pilots (Endsley 1999; Garland et al. 1999). The same should apply to railways.

Examples of expectancies from railways

As was shown, the mental picture in the mind of the observer strongly determines where information is looked for and how this information is interpreted. Repetitive SPAD incidents, for instance in Weesp, Boxtel and Rotterdam give rise to the belief that this also plays a role in railways (Carstens et al. 2007; Inspectie Verkeer en Waterstaat 2007d). In these cases, drivers appear not to have picked up information from which they could have derived that the situation did not match their expectation. Yet the question remains whether this is a driver's fault or a problem of the signal being unable to catch the driver's attention sufficiently (Davis 1966) or being unable to update the driver's mental picture. An example where habituation played a role and a signal clearly did not have any surprise value for the driver was a near miss in the Tilburg West station (Inspectie Verkeer en Waterstaat 2008d). A driver completely overlooked a stop signal, since this signal was normally indicating to drive on sight and he had already seen that the track ahead was clear. However, this time the signal showed a stop aspect. In both cases this signal would be announced by a yellow aspect in the preceding signal, and the driver who noticed the yellow aspect could not identify that another situation was actually occurring than was normally the case. Here, the signalling system did not provide enough information and thus (falsely) confirmed the driver's expectancy. This may also explain why shunting trains have a relatively high number of SPADs (NS Reizigers 2008). As these movements often go from the platforms to the yard or vice versa, they always

have only one clear block ahead (sometimes two, combined with a yellow-yellow-red aspect sequence), which means that drivers of shunting trains see signals at yellow all day long, and a yellow signal will come to mean 'you may now go!' rather than 'prepare to stop'.

Expectancies are based on a mental model, which in turn is based on learning from previous situations that are similar. A driver knowing that a yellow signal aspect becomes upgraded (e.g. green) every time shortly after he has passed it, will thus learn that a yellow aspect is meaningless at that particular spot; the warning that it ought to provide is then in fact no longer given by the yellow signal itself, but by the absence of the movement authority upgrade shortly after passing the signal. This of course is a very implicit way of providing a warning and thus care should be taken that yellow signals aspect are not normally warning for a danger that is hardly ever present in practice. Such situations were widely used in the United Kingdom to cope with a signalling system that has insufficient variety in its distant signal aspects to unequivocally announce the main signal aspect. More concretely speaking, speed reductions at junctions were often announced with a 'prepare to stop' aspect; the main signal then normally cleared (with a speed restriction) after the train had proceeded a certain distance and would thus have slowed down to about the required reduced speed. This is called approach control, but it bears risks in those cases where the main signal will *not* clear upon approach: 'expectation of an approach-controlled signal clearing every time can, in the mind of the driver, generate an unintentional meaning resulting in ambiguous information' (Rail Safety & Standards Board 2003).

In railways, no data are collected to see where and how often a yellow signal aspect is *not* followed by a red aspect in practice. Gathering this information need not to be difficult with current technology (Goverde et al. 2008), and this would allow us to find out where drivers are likely to develop dangerous false expectancies, and signal operation could be changed to prevent that. Developments with automatic route setting are likely to have made things worse rather than improving the situation in this respect (see section 8.2.5). The lesson we learn is that a strong and unambiguous message is necessary to correct or adjust an expectancy when it is false. This was also concluded in road traffic research: 'with timely warnings that something actually changed in the driving environment, an adequate action is likely to result' (Martens and Fox 2007).

Lessons from road traffic research

To clarify the notion of expectancy, we will have a closer look at road traffic. Although the channels through which information is perceived by drivers may differ between the two modes (road traffic operates on sight and rail traffic by guidance of traffic control), both car and train drivers must find information and develop a driving strategy (in fact this is performing a DDA sequence), so that the role of expectancies will probably be similar for both kinds of drivers. Houtenbos (2008), who performed research in road traffic expectancies, states that whether an expectancy was correct or not will only become clear after the situation, when the real outcome of a situation can be compared to the expected one. This seems to be different in railways because a driver's expectancy can be compared to the plan of the rail traffic controller (in road traffic, there is not a central authority with a plan, and thus there is nothing to compare to upfront). Road users have two types of expectancy: a long term (or a priori) expectancy and a short-term (or ad hoc) expectancy (Knapp 1998/1999). Indeed, these two types of expectancy were also recognisable in the train driver interviews. The short-term expectancy could even be further divided into four sub categories, two of which are based on information provided and two which are based on 'default values' (section 7.4.2).

Long term expectancy
An a-priori expectancy is based on experience, upbringing, culture and education. It is by definition a general notion, a scheme, of what kind of events can be expected, without knowing precisely what exactly will happen at what time in the current situation. An example from road traffic is the expectancy that there is no oncoming traffic if driving on a motorway. From interviews with train drivers the following examples have been found:

- Route knowledge: the ability to know where along the route additional attention is required;
- Rolling stock knowledge: the ability to estimate rolling stock behaviour, especially braking characteristics;
- Understanding the principles of train protection systems: the ability to understand which guarantees are given where by the signalling system and ATB system;
- Knowledge of the timetable: the ability to expect whether a slow train is ahead so that speeding up is not useful; or knowing where to start coasting to enable an energy-efficient driving style.

Short term expectancy based on provided information
A short-term expectancy is based on local practices during a particular trip, such as actual information that is perceived by the driver:

- (By definition) meant to further detail the a-priori expectancy
  Drivers did not give many examples of this category, possibly because it is an activity at a skill-based level – asking how a driver adjusts his expectancy then is a question similar to asking how someone pronounces an 'A'. It can be assumed, though, that a driver knows from his route knowledge where he can be led to a different track and then uses the information provided in lineside signalling to identify whether that will happen this time and, if so, what is a correct speed to change tracks with. Some drivers indicated that they deliberately tried to forget previous situations in order to avoid wrong expectations; so they force themselves to always expect a red signal aspect unless they notice that the real aspect is better.
- As a reference for the real situation
  Nearly half of the drivers indicated that they expect a certain aspect in a signal and that they check whether they indeed see that aspect in the signal. 'If it does not correspond to the expectancy, you will immediately notice that'. This is in line with earlier research by Branton (1979) and the conclusions of Hale and Glendon mentioned before. The long term expectancy of the signal aspect is then being checked against the short term information. However, it requires the signalling system to indicate all different scenarios with a different aspect. If this is not the case such as in the Tilburg West example mentioned above, where a stop before the station was announced with the same yellow aspect as what is normally the case when the train can enter the station, incidents are likely to occur: the driver is not sufficiently made aware of the changed situation and the information perceived then tends to confirm the long-term expectancy. Also the automatic route setting system ARI has its effects on the train driver's expectancies. One driver explained: *'if you approach a yellow signal, you will often also know where the trigger point for ARI is [which is long term expectancy or route knowledge] and thus you expect when the signal will become green. If that does not happen, then you are in some way waiting for the signal to be cleared. If you then pass it and it is still showing the yellow aspect, this may still come as a surprise, even though you could see the signal for a very long time. Now we have to adapt ourselves to the system, but this should be the other way around'*. This is a difficulty in adjusting the short-term expectancy because the information of the yellow aspect is ambiguous before the ARI trigger point is passed: it is not clear whether the route

is just 'not yet set' or whether it cannot be set this time. In terms of Ashby's Law: the yellow aspect is used for two messages and thus has too low a variety.

Short term expectancy based on 'default values'

- To complete unavailable information
  As explained in Figure 7.5 a number of situations exist in which the same signal can clear routes with different lengths. In all cases the signal aspect will be yellow. Again, several different messages are packed into the same signal aspect. The driver has to regulate his train's speed either on the basis of an assumption of the route he will be provided with, or he has to define a speed that is low enough to fit every possible route. The latter however, is a far more difficult task than the normal task of reading the signal and acting, since he must not only mentally calculate the most restrictive speed profile for one route, but for all possible routes. Moreover, this leads to an unnecessary low speed in most cases. Many drivers indicate that they lack information about the traffic controller's intention and would like to know how they can adhere to the traffic controller's plan. Here drivers are being forced to rely on long term expectancy, rather than giving them access to the short term expectancy the controller has set up for them. This clearly is a communication problem.
- To replace information
  In some cases, a short term expectancy is used to replace information that is actually provided. Examples are relying on the ATB cab signalling instead of the lineside signals (although ATB cannot display all lineside signal aspects and thus only provides limited information), or routinely braking to a speed that is normally correct in a certain situation, although now a lower speed was indicated.

No clear criterion was found for when a long-term expectancy is being changed into a short-term one. This is dependent on when information becomes available to indicate that the long-term expectancy can be confirmed or denied, thus becoming a short-term expectancy.

Forgotten knowledge: don't rely on the signalling system only

I made a remarkable find when reading an old education document which was produced by the Dutch railways (N.V. Nederlandsche Spoorwegen ±1952). The exact publication date is unknown, but given the signalling systems that the document mentions, it must be dated between 1946 and 1954. The document in fact made clear that some problems with expectancies had already been solved in the past: with regard to the yellow-yellow-red sequences (in the Netherlands nearly always used along platforms that are divided so that two trains can stop behind each other, see Figure 7.5), similar situations existed, albeit with mechanical signals. In these situations, it was only allowed to let a train stop halfway along the platform if it had a scheduled stop at the station. For shunting movements, the same document indicated that it was necessary not only to grant the driver a movement authority by clearing a signal, but that he should be told additionally to which track he had to run and how far.

Both are examples in which the expectancies of the driver are clearly taken into account, and in both cases the signalling system was not the only thing relied on to provide a safe route. Since such additional rules appear to have disappeared, this has to be cured by the driver by adopting a careful driving strategy. This is especially true since the current yellow signal aspect provides less information than it did in those days. Indeed, when the current signalling system was introduced, the driver was required to slow down when he had passed a yellow aspect and then to drive on sight. This requirement is currently not as specific, as will be shown in section 9.4.4.

Conclusions on expectancies
Expectancies develop quickly and help drivers to perform their task efficiently, focussing their attention on what is believed will become important. They are based on a mental model, which in turn has been formed by learning from previous similar situations the driver has experienced. An expectancy can be defined as a situation-specific (dynamic) idea of what is going to happen, based on what is known from:
- a general, static knowledge about the system's behaviour (education, route knowledge);
- earlier experiences with similar situations (habit-forming).
- earlier perceived information on this specific situation (information);

Information that comes unexpectedly and is in any way ambiguous has a high probability of not being noticed or, if it is noticed, of being distorted. However, if the information is able to attract the driver's attention and the message is sufficiently unambiguous (so that it cannot be distorted), it will have a very high conspicuousness. If the driver is on the look out for such anomalies, then route knowledge and long experience helps him to pinpoint them. In other words: information about deviations from an expected situation must be 'pushed through the driver's expectancy' in order to correct it, or 'dragged through' by the driver himself, being alert for change and expecting the unexpected.

Especially because signal aspects are used to confirm an expectancy it is of utmost importance that signal aspects make a clear difference between different scenarios – exactly what Ashby's Law says on a more general level. A simple and formally correct rule such as 'a yellow signal aspect means that the next signal can be red' may be rendered ineffective if in practice the next signal in the vast majority of the cases is *not* red. Drivers will learn that this yellow is usually unimportant information which should not be paid too much attention to. This may be difficult to accept for people with a technical background. However, it is not realistic to think that drivers should or could abandon their expectancies. If in an incident a driver missed a red signal, we often read in the report that the driver 'had an expectancy', and it sounds as if this were the cause of the SPAD. However, a driver will always have an expectation, and he needs that, too. The correct conclusion should be that the driver had a *false* expectancy and thus was not aware which element in his environment was the most important one to focus his attention on. This means that if an accident is attributed to 'driver expectancy', it should also be investigated how each of the three elements mentioned (education, information and habit-forming) of which the expectancy consists influenced the course of events that eventually led to the false expectancy, the false prioritisation of attention and eventually to the SPAD. The objective is to make sure that the driver's expectancy, which is always there, does match the situation as intended by the traffic controller. This requires exchange of information, or, if this is not possible, it requires that the driver realises all possible situations that may occur. The long-term expectancy can be changed by education or (simulator) training, so that the driver knows a broader repertoire of possible situations to expect on the short term. Short term expectancies can be changed by providing information about the actual situation.

The next section deals with distraction, a concept which is linked to expectancies, as will be made clear.

7.4.3 Driver distraction

Ill-defined term
According to official statistics (Inspectie Verkeer en Waterstaat 2007f), driver distraction is an important cause of SPADs. It was considered by the inspectors to be the cause of 17% of

all SPADs in the period 2002 – 2006. But what is it? The report defines distraction as 'decreased attention'. In this report it is also mentioned that the distinction between causes such as 'distraction' or 'expectancy' may be subject to the individual interpretation of these words by the driver who happened to be involved in a SPAD. We will see that these two concepts indeed are linked, but also that having an expectancy is not as bad as is often thought, as was also shown in the previous section.

Both terms, expectancy and distraction, are not clearly defined, which is not unique to railways. According to Sheridan (2004), also in road traffic there is considerable vagueness about the exact differences between frequently used terms such as attention, alertness, vigilance, focus and situation awareness. Sheridan defines driver distraction as 'a process or a condition that draws away driver attention, thereby disturbing driving control.' He proposed to make the concept of distraction clear by describing it in terms of control processes, an approach that is especially useful for this thesis.

Distribution of attention

The problem with distraction is generally that several tasks require some of the operator's sensory, cognitive and motor resources – the same three elements of which also the DDA sequence consists. Such different tasks can both be intended (looking at signals and looking at the timetable) or one of them can be unintended (looking at signals and looking at a road accident alongside the track). Thus, the driver needs to distribute his attention. Sheridan states that the frequency of sampling for two dynamic processes should be proportional to the bandwidth (maximum rate of change) of the processes; in other words, if in a process things only change slowly, it is not necessary to identify the current state very often. This is confirmed by a number of studies; see Sheridan (2004) or van Westrenen (1999). Thus, if a driver does not expect a change in one of the processes under his control, he can afford to devote attention to other tasks. A problem arises if unexpected events may happen in one or both processes, as this makes the rate of change unpredictable. Therefore, it cannot be said which of the two processes will need attention first. Indeed, also research in a railway context shows that attention conflicts can be caused if there are several information sources (Jansson et al. 2005). The problem can unfold itself in a number of ways:

- When busy with one task, a cue is unintentionally missed in a more important task, thus making sampling risky (Sheridan 2004).
- When busy with an important task, a cue presents itself in a less important task. According to Sheridan stimuli that are sudden and unexpected cannot be intentionally ignored, even if they are irrelevant to the current task. One of the interviewed drivers gave the example that ringing of his cell phone (which can be considered sudden and unexpected) 'captures' his attention and thus distracts him from the driving task, even if he decides not to pick up the phone – it is still a decision that he must make.
- If an irrelevant stimulus is expected, it can be intentionally ignored (Sheridan 2004). This may be one of the mechanisms that allegedly cause the ATB cancel function to be ineffective (see section 9.4.2).
- Finally, if a person is (almost) fully occupied with one task, he may miss information even if it is clearly presented. This is what was described in the previous section as inattentional blindness.

Information filters or other equipment may assist the driver in determining which information needs his attention first. However, one problem with providing such tools to attract the driver's attention is that the driver may become complacent, which means that he relies on the system to warn him rather than being attentive himself. It is one of the ways in which the risk homeostasis theory (Wilde 1994) manifests itself, since such tools may make the operator's

work easier. Some drivers are reported to have adopted such a driving style in relation to the ATB warning bell. Refer to section 9.4.1 for more information.

Conclusion on distraction
Distraction in fact means that a driver did not pay attention to the most important element in his environment. This means he needs assistance to prioritise tasks. Unexpected high-priority events therefore need to be accompanied by an alarm, whereas low-priority events should not distract the driver's attention from whatever he is doing. Especially as long as no such prioritisation tools are provided to help the driver with that, an expectation is a necessary prerequisite in order for the driver to get an idea of what is going to happen and what is important. Also in this case, an expectation is not something undesired, but something that helps the driver to distinguish between relevant and non-relevant information, and to adopt an anticipatory driving style.

Where the railways see expectancies and distraction generally as problematic issues, two other human factors issues are regarded as positive properties: these are route knowledge and vigilance, the topics of the next two sections.

7.4.4 Route knowledge

A well-known term for an uncertain concept
Route knowledge is a characteristic term in railway operation and is probably as old as the railways themselves. In the steam age, route knowledge helped the driver to find ill-lit gaslamp signals by looking through a small window in a noisy cab whilst carrying out many other tasks to keep the engine going. Noticing the signal was extremely important as there was no backup of a train protection system. Now, with bright colour light signals, well designed and often air conditioned cabs with large windscreens and more and more driving tasks being automated, and with the backup of an automatic train protection system, one would expect a decreasing role for route knowledge. Dutch legislation (*Besluit Spoorweg-personeel* 2004) requires train drivers to have route knowledge, where it is the responsibility of the train operating companies to define the content and how they will examine it. This is not subject to any specific requirements (Inspectie Verkeer en Waterstaat 2009a). Neither from my interviews nor from literature, such as education material or professional literature on signalling, or Dutch legislation did it become clear what route knowledge exactly is and why it is necessary, a remark that was also found by the VSD, a Dutch association of new train operating companies. Also a recent British study into railway's route knowledge had difficulties to define it (Air affairs 2006).

What is route knowledge?
Although nobody appears to know what exactly route knowledge is, people generally agree that if a driver does not have it, he is likely to get into trouble. Drivers themselves attach much importance to route knowledge. Having route knowledge about the entire network or large parts of it is generally seen as a great achievement for a driver, and it is seen by many of them as a necessity to master difficult station layouts. One driver said: 'a good driver does not run through a red signal at that location; he would know that that particular signal is poorly visible. That is what you have route knowledge for.'

Basically, it is more reliable to retrieve information from memory when there is a clue that assists in the retrieval (Encyclopædia Britannica Online 23 Jul. 2009b). Such a clue is not there if route knowledge is the only way to correctly respond in anticipation to a signal. Besides the fact that it is questionable in the first place to rely solely on a fallible human memory in order to be able to stop in time if a signal is poorly visible, it became also clear

from interviews that drivers have considerably different ideas about what route knowledge is. Some drivers stick to a quite general notion of knowing names of stations and junctions, positions of difficult signals et cetera. Others claim to have a far more detailed understanding, which consists of knowing the location of all 10,000 signals in the Netherlands, or even the interdependencies in their aspects and all possible aspects they can show, and knowing the length of passing tracks. Some drivers told that they start braking for a certain station when they pass certain landmarks, e.g. the local supermarket. Although this does work, the existence of this knowledge and a clear view on it are not controlled by the railways.

While this research was being carried out, route knowledge was finally defined by the European Commission in the Technical Specifications for Interoperability (TSI) 'Traffic operation and management' (European Commission 2006). The definition is given below:
'*Route knowledge comprises the specific knowledge and/or experience of the individual characteristics of a route that a driver needs to have before being permitted to drive a train over it on his own responsibility. It comprises the knowledge that is needed in addition to information given by signals and documents such as timetables and other on board documents and in addition to knowledge of operations and safety rules applying to the route and as specified in clause 2.2 of this Annex. Route knowledge covers in particular:*
- *The conditions of operation such as: signalling and control and communications*
- *Knowledge of the location of signals, steep gradients, and level crossings*
- *Points of transition between different operating systems or power supplies*
- *The type of traction power supply on the line concerned including location of neutral sections*
- *Local operating and emergency arrangements*
- *Stations and stopping points*
- *Local installations (depots, sidings,...) as required by the company.'*

Although this definition is pretty much in line with common practices, it still does not tell what exactly the function of route knowledge is or why we need it.

Functions of route knowledge
From my interviews, the following main functions of route knowledge were identified:
- helping the driver to identify the train's position;
- help to expect what will come up along the route, which leads to less stress and thus a more relaxed driving style (better anticipation, better mental model);
- knowing where 'difficult signals' are located (compensating signalling design problems).

A possibly workable starting point to define the function(s) of route knowledge stems from British research: "*route knowledge is more than the information held within a driver's memory; it is the skilful use of this information to interpret the current and future state of the environment. Therefore, any model of the psychological processes involved in route knowledge should incorporate psychological constructs such as situation awareness and attention as well as considering issues relating to memory. (…) these constructs have not yet been widely studied within the railway industry*" (Luther et al. 2007).

Maintaining route knowledge
From interview results it turned out that both the practical and the formal validity of route knowledge are problematic. Several practical issues were mentioned:
- Drivers do not always have memorable points along the route so that they do not know where they actually are, especially if the view on orientation points is restricted, such as in darkness, in fog or when noise screens are installed next to the track.

- When it comes to maintaining route knowledge 'on the fly' some drivers complain that they are always sent via the same tracks through station areas and can therefore not see whether (or how) things have changed on other tracks.
- Also, many drivers reported problems in receiving updates of route information, for instance when renewed lines were reopened. Reasons for this were identified both in late provision of updated route information from ProRail to train operating companies, but also in poor accessibility of the information within train operating companies, such as only during office hours. According to managers, these issues have now been solved.
- One driver warned that too much route knowledge in combination with poor updates could also lead to safety issues, for instance when a driver knows from route knowledge that there is *no* signal at a certain location, he may not look for one and miss it if a signal has been moved to that location. Indeed, this mechanism led to a SPAD when a signal's red lamp failed in a newly relocated and poorly visible signal in a tunnel (Inspectie Verkeer en Waterstaat 2008c).

The examination of route knowledge differs between train operating companies. Some cases were reported in the interviews in which the exam was not more than asking the question 'have you been there?' and answering 'yes'. In other cases, the exam is more thorough, consisting of questions as 'what should you be careful for on track x in station y?'. It remains questionable whether the examination of route knowledge is sufficiently reliable – not only because wide differences exist, but also because railway companies are not clear what function route knowledge precisely should fulfil in the driver's task.

Excuse for unclear signalling layout
Not having route knowledge was identified as a main cause for a SPAD accident that took place in Arnhem in 2006 (Inspectie Verkeer en Waterstaat 2007c). However, a further analysis learned that there were also a number of flaws in the signal layout which were confusing the driver, such as the confusing yellow-yellow-red signal aspect sequence, in which the red signal was hardly identifiable against its disordered background. Also, there were more tracks than signals (Van den Top and Groeneweg 2007). These were identified as difficult situations in sections 7.3.2 and 7.3.4. Also, the train was slightly delayed and the driver did not know the traffic controller's plan. Although this did not result in reckless driving to make up time, it may have created uncertainty for the driver and as such can be considered a contributing factor.

Also from my interviews, it appears that when route knowledge is used as an argument, this often indicates a situation in which the signalling layout is difficult. This is hardly fair to the drivers, as it shifts the onus onto them for recovering the poor decisions of others. Apart from that, route knowledge is a static representation of the situation, learned by previous experiences of the driver at the same location, and it can therefore not be used to fill in any absent dynamic information that differs from case to case. Furthermore, if different routes require different behaviour from the driver (as in the example given in section 7.3.3), route knowledge will not be very useful if the driver is not informed about which route has been set for his train, or if he cannot determine which of the visible signals apply to his track.

Route knowledge: useful and necessary, but do not rely on it for safety
Although no agreed function has been found which route knowledge serves, its necessity under current practices cannot be denied. It would be worthwhile to find out which functions route knowledge is supposed to have and then to consider whether route knowledge is also the best possible way to fulfil these functions. As British research shows, a distinction must be made between novice and experienced drivers, since experience changes the way in which

drivers use lineside signs. Novices need to read the signs whilst experienced drivers use signs just as a confirmation of what they already knew (Farrington-Darby et al. 2005).

Currently, route knowledge appears to be considered essential to be able to drive safely. The question is whether that should be the case. It is primarily the task of the signalling system to supply information on safe driving behaviour. It appears that route knowledge is now often used to compensate for shortcomings in the signalling system's design (signal aspects and their meanings) and application (actual signal configuration). However, even the best route knowledge is useless when the driver is sent to a different route than he expected and the signalling system did not make this clear. If route knowledge serves to mitigate poor signal visibility, it forms an essential part of the safety system and should be formalised more than it has been now, but even having a proper definition will not support the driver in finding poorly visible signals. Route knowledge should therefore not be the last and only measure to ensure the safety of rail traffic, which it now is in some cases. To improve that, the design problems in the signalling system and in the signal layout should be solved, either by changing the signals themselves or by introducing some form of cab signalling.

Route knowledge can be considered, however, as a basic familiarity with the route, and may as such serve one of the conditions to achieve the lowest human error rate. As such, it is not different from having a long term expectation, as was described in section 7.4.2. Since both route knowledge and expectancies are the result of experience, it seems to be impossible to require drivers to have outstanding route knowledge but to expect them at the same time not to have any expectancy about the situation.

7.4.5 Vigilance

Route knowledge helps drivers to focus their attention to those places where information is to be expected. Another approach is driver vigilance, which tries to make sure that drivers can detect information even when it comes unexpectedly. The term vigilance is closely related to other words with similar meanings, such as alertness, attention, or focus, so that its meaning is not always precisely clear (Sheridan 2004). The Encyclopaedia Britannica says the following about vigilance: *"Vigilance refers to the state in which attention must be maintained over time. Often this is to be found in some form of 'watchkeeping' activity when an observer, or listener, must continuously monitor a situation in which significant, but usually infrequent and unpredictable, events may occur. Vigilance is difficult to sustain. No single theory explains vigilance satisfactorily, probably because of its complexity. In the first place, there is a distinction between sustaining attention in a detection task, where the overall workload is high, and sustaining it when little is happening except for the occasional looked-for events. Under both conditions performance can decline over time. During the course of a watch, expectancies develop about the frequency with which [relevant events occur]. If [an event] occurs after an atypical interval, it is less likely to be detected"* (Encyclopædia Britannica Online 23 Jul. 2009a).

The level of attention, and thus also vigilance, can be higher or lower from moment to moment, depending on several factors such as a noisy environment and sleepiness, or task load (under a too high or too low task load attention will decrease). Although monotony can be detrimental to performance, this is not always the case, as recent research indicates: '*monotony-related effects appear only when the cognitive task elements are simple, hence they do not involve just repetition of perceptual-motor task elements. It seems that monotony effects do not appear in cognitively complex tasks, even after prolonged task durations*' (Dunn and Williamson 2009). What is worrying in this respect is that automation of several parts of the driving task has made the driver's task less demanding and therefore more prone

to vigilance decrement; indeed, a computer simulator experiment showed that the ability to detect an unexpected event severely decreased when trains were running on auto-pilot, whilst it was highest with in-cab signalling (Spring et al. 2009). It would therefore be useful to see whether additional task elements can be introduced that keep the worker cognitively engaged in the task (Dunn and Williamson 2009).

Thus, the concepts of route knowledge and vigilance try to achieve improvements in driver behaviour with opposite mechanisms – one improving the expectancies of drivers where information can be found, the other making sure that unexpected information will be found as well. Despite the opposite approaches, both solutions suffer the same imperfections: very high levels of vigilance are difficult to obtain, just as it is difficult to have a mental road map of the whole network. So, vigilance and route knowledge are useful and important, but they should not be relied on as the only measures to prevent a SPAD.

Since it was expected that better vigilance would reduce the number of SPADs, the SPAD workgroup decided to introduce a vigilance test for all drivers. In terms of our control model, a vigilant driver is a prerequisite for performing a DDA loop: someone who is not vigilant will have difficulties to detect relevant information. Yet when a DDA loop fails, one cannot automatically say that the driver vigilance is insufficient: the problematic situations described in section 7.3 generally refer to diagnosing problems or to the inability to detect the signal in time because it is visible only at a late moment, or it is not immediately clear to which track it applies (the latter being a problem that can be both a detection and diagnosis problem).

7.4.6 Conclusion on driver human factors

In railways it is often said that drivers should not have expectancies, but that they should have route knowledge. Maybe this can be explained because route knowledge is what drivers are taught and therefore must be good, whereas expectancies are only found in incident investigations (because no one looks at expectancies when there was no incident) and thus must be bad. However, route knowledge can be seen as a special case of expectancies; both are formed with increasing experience with a situation. It is therefore not possible to require drivers to have route knowledge and simultaneously not to have expectancies, even apart from the fact that having an expectancy is a normal part of how humans works.

Another common train of thought for at least some people in railways is that a driver should not have expectations but should simply follow the orders of the signals precisely. However, at least two issues remain unsolved in doing that. One problem is the ambiguous signalling information (e.g. yellow-yellow-red sequences), where the signalling system's own logic is bypassed; the other one is that the speed signalling system cannot indicate the point up to which a route is set. In order to provide the driver with a correct mental picture of the situation he is working in, it is required to have sufficient unique signal aspects; 'intelligent tricks' with signals to avoid having to indicate additional aspects will sooner or later lead to ambiguous information provided by these signals.

Another important lesson is that expectancies govern where attention is focussed to. This must correspond to the prioritisation of tasks the driver has to perform. If information with a high surprise value is presented in an unimportant task, it may still capture the driver's attention and thus distract him. Other operators should therefore not distract a driver (e.g by calling him) when he is in a critical task, such as approaching a red signal. A possibility to prevent such distractions is to provide information filters, e.g. so that the phone will not ring when a yellow signal aspect has been passed. Likewise, essential, unexpected information in

important tasks (e.g. the message to prepare to stop) must have a high surprise value and be unequivocally clear if it is to attract attention.

Because attention is paid to those tasks where information is expected, it is difficult to maintain vigilance: this is paying attention to a task in which it is difficult to expect when and where information will be presented. The importance given to the vigilance of drivers indicates that the railway's information system is not tailored to the human operator working in it: he must adapt to the system and 'catch' information when it is provided;

It would be useful to use situation awareness as a framework to understand expectancies, distraction, route knowledge and vigilance. A main issue from situation awareness literature is that information should be provided in an integrated way and such that it enables drivers to think ahead, which was also found in a railway context (Jansson et al. 2005). It would be advantageous if the operator could decide, depending on task load, when to pay attention to which information sources. For example, cab signalling would relieve most of the mentioned problems:

- It solves problems with poor signal visibility or relating the signal to the correct track;
- It is readable whenever the driver wants to do so;
- It can provide a far higher number of messages than a lineside signalling system can;
- An information filter could be coupled to it, so as to avoid distraction.

7.5 Train control in safe envelope terms

7.5.1 Overview of information provided

The previous section on human factors showed how drivers, or operators in general, form a mental picture of the situation using DDA loops. Which are the relevant information elements the drivers should represent in their minds? In section 4.9 a number of such elements were derived from the safe envelope concept. Table 7.6 provides an overview of the extent to which current practices fulfil these requirements. Broadly speaking, the train driver must respect two operational envelopes; one concerns efficient infrastructure usage and stems from the functions of timetable design and traffic management in the cascade model; the other one concerns infrastructural limitations that stem from the processes of infrastructure design and infrastructure allocation.

7.5.2 Remarks on the controllability envelope

In a calculation made by track designers, the last point of intervention in order to remain within the operational envelope is calculated and shown with trackside signs and signals. Since these trackside signs do not know the train's properties, these warnings are necessarily based on the train with the highest speed and/or the worst braking characteristics. After passing that point, the railways rely fully on driver experience to handle the train, based on route knowledge (e.g. braking less or even giving traction when drivers know that an uphill gradient will slow down their train) and knowledge of the train's braking characteristics. Several drivers indicated that for a certain rolling stock type the braking performance varies quite a bit depending on the actual unit they are driving, which makes performance unpredictable especially in the beginning of a trip with that train. Here again, it is clear how important it is to be able to make a good estimation of the system's future behaviour. Given the fact that the train driver is ultimately the person who can control the train's kinetic energy, the aim should be that his situation awareness is improved wherever possible: without a correct mental picture of what is going on, it is not possible to take the right decisions.

Table 7.6: Overview of current fulfilment of safe envelope requirements for train control.

Safe envelope requirement	Traffic management (time/spatial domain)	Infrastructure allocation (spatial/physical domain)
Actual state	Driver knows where he is, but he can only see whether he is on his path at discrete points for which a timetable entry exists	Available at discrete points (lineside signals)
Goal state	Available at discrete points for which a timetable entry exists, but not up to date	Not applicable
Operational envelope	No, the plan contains no explicit tolerances	Yes, movement authority given by main signals
Attainability envelope	Not provided (e.g. amount of running time margin)	Yes, implicitly provided by signalling distances which allow for attainable braking distances
Controllability envelope	Based on driver experience (e.g. slopes in track, slack in timetable and braking characteristics of train)	Last moment of intervention indicated by distant signalling and by driver experience (e.g. route knowledge)
Viable envelope	Not applicable; damage can only result from physical interaction, not directly from planning.	No need for this when the train is supposed to remain within the operational envelope. Once outside operational envelope, no information is explicitly provided, but some drivers manage to stop on sight even when not explicitly instructed to do so. Relies on driver when driving on sight

7.5.3 Remarks on the viable envelope

Since the railways operate on the principle that infrastructure is assigned to one train at a time and that trains do not leave their territory, the viable envelope (consisting of all things the train could bump into or could otherwise exchange energy with) is not explicitly marked. However, it must then be made clear to drivers as well as traffic controllers when this principle is violated for whatever reason. This is currently not the case: once the process runs out of the infrastructural operational envelope, for instance after passing a signal at danger, a driver hardly has any information which can tell him so. In some cases drivers notice it when points are not set correctly or when other trains are in clear sight heading for the same track, but cases are also known where a driver only became aware of a SPAD when he arrived in the next station (Inspectie Verkeer en Waterstaat 2005b).

7.5.4 Usefulness of models

In summary, the theory developed in Chapters 4 and 5 was helpful to identify explanations for occurring problems. The safe envelope provides a framework of items on which the driver needs to receive information, as given in section 4.9; the DDA sequence was helpful to analyse whether the driver can acquire the required situational awareness to perform his job adequately. Thus, it was possible to identify situations in which drivers lack some information and situations in which the information is difficult to interpret correctly. The cascade model can be used to see which functions could provide such missing information. This model was

introduced in section 2.3.1; an example of integrated information flows between several actors will be provided in Chapter 10, after we have analysed the information need of traffic controllers in Chapter 8 and some design and management factors in Chapter 9.

7.6 Conclusion

Up to date planning information hardly present
This chapter served to answer the question as to which problems and related solutions for SPADs and delays can be found in the train control process. Passing signals at danger is not just a human factors or technical issue, but also an information exchange issue between traffic controllers reworking the plan and drivers executing it. The signalling system provides some of the necessary information exchange between these two actors, but not all of it: despite the central role for time, drivers can currently not see in sufficient detail whether the timing of their train adheres to the (possibly rescheduled) plan of the traffic control centre; only when a deviation from the desired timing would have consequences for safety do signalling systems inform drivers about this. Drivers cannot always produce the right actions because the variety in the information they receive is too low.

Since drivers do not receive information about the actual plan or the actual traffic situation, they behave mostly on the basis of individual preferences and individual experience with earlier situations. Both do not necessarily lead to behaviour that corresponds to the traffic controller's plan – if there is one, because we will see in the next chapter that this is also not always the case. Since essential information is missing to control the process properly, simultaneous infrastructure claims will result and the signalling system, intended as a backup system, is needed unusually frequently to ensure safe operations. The driver's task can be compared to that of a tightrope walker who is forced to walk in the safety net instead of on the rope. This leads neither to a very exciting performance, nor does it improve safety, which is still at risk if the safety net breaks. When deviations from the intended timetable occur, traffic controllers generally assume that drivers are fully informed by the signalling system (see section 8.2). This chapter has shown that this assumption should not be relied on too easily. The signalling system was intended to control the exceptions (e.g. to prevent that route setting errors and planning errors result in accidents) rather than to control the normal situation as if a railway signal were a traffic light. This means that safety improvements can be gained if the planning and rescheduling processes are redesigned in such a way that fewer exceptions are left at the driver level, and hence that he will be running on green signal aspects most of the time.

In the next chapter, the process run by the traffic controllers will be explained. We will see that traffic controllers suffer similar problems as drivers, because they are not optimally supported in their task.

Signalling: at the same time too simple and too difficult
The Dutch signalling system is easy to understand for drivers, especially between stations, through junctions and in small stations. In more complex station areas, there are a number of weak points where drivers can have difficulties determining the most relevant speed profile. Although these situations are often classified as 'driver errors' or 'driver expectancies', this is a one-sided consideration, since it is a task of the signalling system to inform the driver about what is most important and what he should expect. If a large number of drivers encounter problems in these situations, we must conclude that there is a flaw in the design of the signalling system, although not all of these issues could have been foreseen given the

scientific knowledge of the 1950s when it was developed. The problems which we have revealed are threefold:

1. Yellow, the most important signal aspect, is used to announce too many different situations. The driver is not prepared optimally for a stop, because he only knows that he has to stop, but not the exact location at which to stop. This makes a clear diagnosis of the situation difficult, especially in those cases where drivers always encounter yellow signal aspects, but in different situations.
2. A driver must be able to reliably relate a stop signal seen at some distance to his track, but it has been shown that this is not always possible.
3. The sighting distance of the stop signal itself is in some cases too small, so that the driver does not have enough time to respond and stop short of the signal. In some cases this is mitigated by a driver's route knowledge, although this should not be relied on. Even in situations complying with ProRail's design rules, the sighting distance can be too small for the driver to react.

From incident reports and the SPAD workgroup database, it appears that SPAD investigations often consider signals in isolation and not in relation to the sequence of preceding aspects in which drivers encounter them, although these are essential to correctly build up the driver's situation awareness. A positive exception are the SPAD workgroup's investigations of repetitive SPAD locations, which have often resulted indeed in measures taken in the preceding signal sequence. Also in the inspectorate's database of SPAD accidents, the correct perception of the preceding warning signal has only been registered since 2005 (Inspectie Verkeer en Waterstaat 2007f). This is especially striking, since missing the preceding signal (and thus not being prepared to stop) was the main cause of the severest train accident in the Netherlands, which took place near Harmelen in 1962[9], so that one would expect that the importance of distant signalling would be readily known.

Often it is said that SPAD accidents find their cause in the fact that the ATB system does not function below 40 km/h. We have also identified that not only the automatic train protection system, but also the signalling system can be ambiguous when indicating speeds of 40 km/h or lower. Both protective systems thus have their weaknesses in the same situations, and the system then relies on the driver to provide a safe outcome (using route knowledge or expectations).

Human factors

Many SPAD incidents (36% of all registered SPADs in 2006) are attributed to drivers having an expectancy or being distracted, but the two concepts are linked and not clearly understood and defined – at least not in the context of Dutch railways. A possible cause for this may be the railway's engineering culture. In any case, it is striking that at least one third of the incidents are attributed to mechanisms that are not well understood. It was shown that having an expectancy is not a human weakness that leads to accidents, but a normal behaviour that enables human decision makers to guide their attention and to focus on those things that are expected to be important, which in turn leads to efficient behaviour – at least as long as the expectancies are right. It was shown that expectancies build up quickly; in road traffic it is a matter of days rather than of months or even years, as is sometimes thought in rail traffic.

Both false expectations and distraction point towards insufficient situation awareness. The question should therefore not be how to get rid of driver's expectancies, but how to prevent

[9] This accident also led to installation of the automatic train protection ATB, which made it impossible (in most cases) to let a warning signal pass unnoticed and without braking.

that they can become wrong. DDA loops should be in place to either confirm or replace a driver's expectancy with up to date information. Also, a clear definition and categorisation of expectancies should be introduced. At the moment the term 'driver expectancy' sounds interesting but explains little. Also route knowledge is in fact nothing different than a long-term expectation, i.e. one that is not specific for the current situation. Since both route knowledge and expectancies are based on experience, it is impossible to require drivers to have outstanding route knowledge and simultaneously no expectancies. Another peculiarity with regard to route knowledge is that everybody agrees that it is important to have, but no one has defined why exactly it is necessary and a precise formulation of what it should consist of has only been in place for the last few years. It often appears that route knowledge is put forward to justify the presence of difficult signalling layouts.

8 The traffic control process

8.1 Introduction

In the previous chapter, the process of controlling a train and the problems that arise therein have been explained. This chapter deals with the current process of traffic control. The role of traffic control is to (re)schedule infrastructure use such that trains have conflict free paths as much as possible, and to set routes accordingly. By doing so, traffic control influences the operational ranges for time, speed and position for train drivers. A plan which is not conflict-free is a prerequisite for both SPADs and knock-on delays. This chapter will explain how ProRail's traffic control organisation tries to keep the plan conflict-free and to set routes accordingly.

This chapter has the following structure:
- Section 8.2 will elucidate the process of route setting.
- Section 8.3 will handle the topic of rescheduling infrastructure use to avoid conflicts, so that routes can be set such that drivers do not need to make unplanned stops.
- Since there are several indications that the network and traffic controllers have poor situation awareness and suffer automation surprises, section 8.4 will discuss some human factors related problems and their solutions.
- Section 8.5 will discuss how traffic controllers cooperate with other actors, which leads to several paradoxes that need to be understood before improvements can be achieved.
- The information found in this chapter will be mapped onto the safe envelope model in order to see whether current practices fulfil the requirements this model states and if not, what has to change. This will be done in section 8.6.
- Finally, conclusions will be drawn in section 8.7.

The information in this chapter comes from the literature, discussions with those concerned at all levels in the rail system, but above all from the interviews with traffic controllers (and to a lesser extent with train drivers) conducted for this thesis; see Chapter 6 on the methodology. Chapter 9 will then deal with the non-operational control processes, such as design and management.

8.2 Route setting: assigning infrastructure to users

8.2.1 Fail safe route setting

Route setting requirements
As was described in Chapter 2, a route is formed when all infrastructure elements (points, bridges, road crossings) are set to the correct position and locked in that position. Then, a message is sent to the train that will be using the route, usually via lineside signals although other media were, are and will be in use. The route elements must remain locked until the train has passed through them; only afterwards can they be used again for another route. In the railway's early days, it was the signaller's task to set all elements manually. Since in some cases routes were set incorrectly, a device was installed that prevents signals from being cleared if not all elements are set to the right position and ensures that they cannot be set again until they are released by the train. This constitutes the interlocking; interlockings were at a later point coupled to track vacancy devices so that signals could only be cleared if the route was set correctly *and* free of other trains.

Fail safe design and its limitations
Analysis of accidents which still happened led to understanding of a great number of failure mechanisms. Modifications in technology or procedures were then carried out to prevent similar accidents happening again (Rolt 1966). One interviewee compared railway safety to DNA: 'what functioned well remained, what did not work was abandoned. Now we have a system that is quite safe, but nobody really knows why'. By and large, interlocking excludes the most common errors during the route setting process from occurring, especially because the interlocking is fail safe. Being 'fail safe is the ability of a technical system to stay in a safe state or to go into a safe state immediately, in the case of a fault' (Deutsches Institut für Normierung e.V. (DIN) 1994). This means, for example, that a signal cannot be cleared as long as the route is not set, even in the case of failures. Because the interlocking and the signal aspects are considered to be a final check on safe operation, these are also indicated as *vital* systems.

However, fail safe designs can only prevent those failure mechanisms for which they were designed. An example of a failure mechanism not covered in system design is the incomplete meaning of the yellow signal aspect (see sections 7.3.3 and 9.4.4). Also, safety may be at risk after maintenance or if parts are insufficiently maintained (e.g. leading to shortcircuiting or to wiring errors; see Jongerius 1993, Inspectie Verkeer en Waterstaat 2004d, 2008a, Railway Gazette International 04 January 2008). Although fail safe interlockings do, in general, protect against errors during the route setting process, and also protect against their own failure, it is too easily believed that nothing can go wrong once a route has been 'accepted' by the interlocking.

Relying on the interlocking for safety is (not) enough
Because of the fail-safe design of the interlocking, traffic controllers are nowadays instructed that they can rely on it unless there are technical failures which require the interlocking to be bypassed (ProRail 2003): also in interviews, they argued that this leads to the correct signal aspects being displayed and that safety will thus be assured as long as the drivers obey their signal aspects. However, even though drivers are very eager to do so, there are certain cases in which this is difficult to do, as Chapter 7 has shown. By the design of the signalling system, the meaning of the yellow aspect is in some cases ambiguous, so that the driver is not fully informed about where he has to stop; in other cases, infrastructure layout is such that it is difficult to determine where the next, red signal will be or whether it applies to the route the train will run over. Indeed, two out of ten traffic controllers said in interviews that they do *not*

only rely on signal aspects when they set certain routes, but that they also inform drivers additionally, since they have had the experience that this could help drivers and thus prevent errors.

Another example of ambiguity is shown by comparing an old educational document about relay interlockings with current practice. It indicated that certain signals were only placed to enhance shunting movements: *'(certain) intermediate signals are meant as shunting signals. Therefore, do not use them right away for stopping train movements.*' Although the Dutch signalling system formally has no separate signals for shunting and for train movements, these signals were not meant to stop a train which was entering or leaving a station (Nederlandse Spoorwegen 1978). Although the interlocking does exactly what it is meant for, which is checking that the route is formed properly (e.g. no points are in the wrong position), it does not know whether a route will be set for a shunting movement or a train run, so that this has to be checked by the traffic controller and/or the automatic route setting system ARI. As one interview with a former traffic controller (and now infrastructure designer) revealed, in those times additional regulations were in place that actually prohibited setting such routes for train movements: however this is no longer so. Not prohibiting this, as is currently the case, may result in surprising locations of stop signals for drivers. It must be concluded that 'safety', at least in this aspect, formerly included the driver's process, whereas it no longer does nowadays.

Although the route setting process is considered to be fail-safe, this only is true for the technologically supported part of this process. Setting the correct route and using it correctly are not fail safe processes. Given the fact that traffic controllers rely on drivers, one would expect that the traffic control system would provide an alarm when a train runs through a red signal. Although such functionality was installed in older versions of VPT, this has been abandoned in newer versions of VPT, which was also commented on in some, but not all, accident investigation reports (Raad voor de Transportveiligheid 2001; Inspectie Verkeer en Waterstaat 2004a).

8.2.2 Formerly: manual route setting on relay interlockings

In order to better understand the current way in which routes for trains are set, it is useful to first describe how this was done until the early 1990s. Modern interlockings not only check the safety conditions mentioned above, but also help the traffic controller to set routes easily. Beginning with relay interlockings, that were introduced from the 1950s, interlockings only required the signaller to push two buttons for every route: one for the starting track and one for the destination track. The interlocking then automatically set the route. Thus the signaller did not need to think of the individual points and signals (means of production) but could concentrate on how to use these optimally given the required train movements (function).

It was common usage to set such routes starting from their far end and working back towards the train; so if the train required a through route from track AE via 21 and 3B to track FV (see Figure 8.1), if possible the part from 3B to FV was set first, then 21 to 3B, then AE to 21. This prevented trains becoming deadlocked, for instance when the last part of the route would not clear due to a failure. Also, it provided the driver at once with the best possible signal aspect (if the route is set piece by piece from the train forwards, the signal aspect would upgrade for each additional step that is cleared).

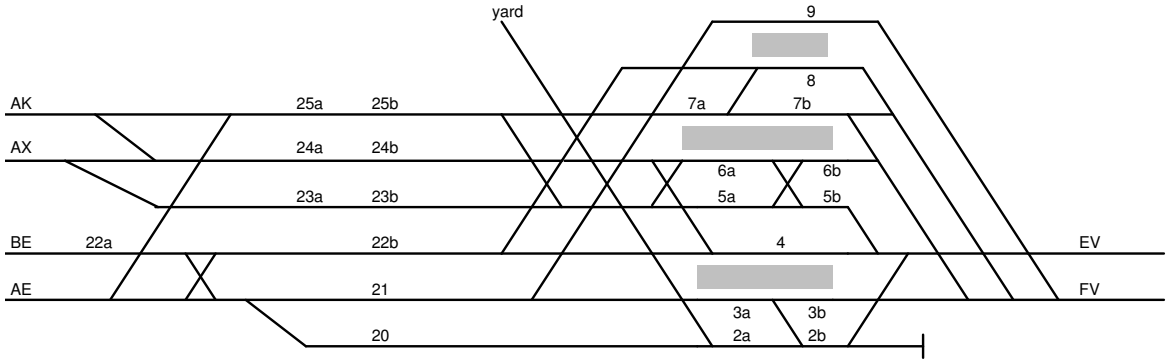


Figure 8.1: Example track layout.

Table 8.1: Example plan lines of the VPT system (23 January 2007, 12:21).

Train number	Movement type	Plan time	Delay	Delay measured at (station abbrev.)	Set time	From track	To track
2039	A1	12:15	+ 12	Ed	12:12	AX	23B
2039	A2	12:15	+ 12	Ed	12:13	23B	4
303037	V3	12:10	V+ 10	Ah	12:19	25A	BE
7539	A2	12:21	+ 0	Ahwa	12:19	23B	8
3040	A2	12:21	+ 0	Ahwa	12:19	21	6B
3743	V2	12:21	+ 0	Ahpr	12:20	22B	BE
124	A	12:23	+ 8	Zvg	12:20	EV	9
7539	V	12:22	+ 0	Ahwa	12:21	8	FV
402039	R	12:23			12:23	4	30

8.2.3 Route setting with VPT

Since the mid 1990s, traffic controllers use the system VPT Procesleiding (VPT Process control, VPT being the information and communication system for timetabling and traffic control). The VPT system contains an automatic route setting module (ARI, see section 2.3.7) and under the name CS/BO, a module for conflict detection and resolution was planned, but never introduced (see section 9.3.4). The information flow in that system is shown in Figure 8.2.

The VPT system uses so-called plan lines which are imported from VPT's timetable design module. Each plan line represents an arrival (A), departure (V), and a through movement (D) of a train, or a shunting movement (R); an example is shown in Table 8.1, which fits on the track layout of Figure 8.1. The first two lines belong to the arrival (A1 and A2) of train 2039. It is scheduled to arrive at 12:15 and will run from track AX via 23B to track 4 (last two columns). The first part of this route must be set ultimately at 12:12 in order for the train not to be slowed down; the second part of the route is due at 12:13. The train currently has a 12 minute delay (+12 in column 4) which was measured at the station of Ede-Wageningen (Ed). Once the train has arrived on track 4, it will become a shunting movement (402039) which will be parked on track number 30 in the yard at 12:23.

If the traffic controller wants to set a route for a train, he can select the plan line on the computer screen and right-click the mouse to execute it. However, normally it is the task of the traffic controller to change these plan lines such that no conflicts will remain, and the automatic route setting system will select and execute them once this is appropriate. Compared to the work a traffic controller did on a relay interlocking, the task is now more that of a short-term planner.

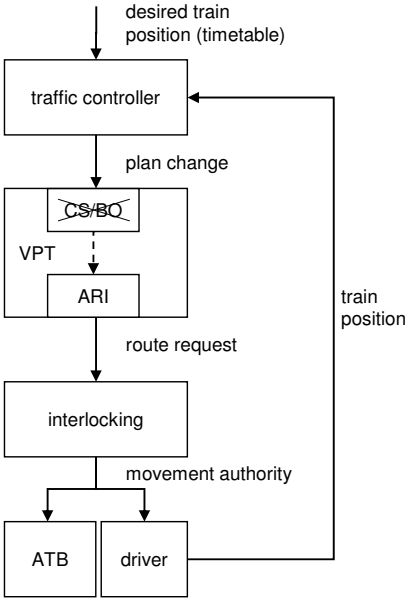


Figure 8.2: Interaction between traffic controller, control system, interlocking and drivers.

8.2.4 Automatic route setting with VPT

Some traffic controllers indicated that they like to work with the automatic route setting system, some said that they had to get used to it and some said that they do not like it. The latter was underpinned with the argument that the traffic controller formerly had to set routes actively, so that as long as he did not push buttons, nothing would change. This was advantageous in the case of sudden disturbances or safety-critical failures for which routes were allowed to be set only after the train driver was contacted by telephone to provide additional information in a written order. Now, sometimes routes continue to be set by the automatic system without the traffic controller being fully aware of this. Also during normal operations, the system may behave unpredictable, albeit it is governed by simple rules. The system will try to set a route either when the departure time of a train is due in one minute, or (for arriving and through trains) as soon as the train has passed a certain location upstream. Although the system is constrained such that it does not change the planned order of trains on tracks, this constraint alone is not sufficient to execute the plan precisely as the traffic controller has it in mind; the system performs the train order check only for destination tracks in a plan line, but not for intermediate tracks or points connecting these. Thus, if two train movements do not go to the same track, the system acts as a 'first come first served' controller and not as one that executes the plan exactly in the way it was planned.

Although these route setting rules are simple, the traffic controller will not always be able to predict which of these mechanisms will be triggered first. In order to remain in control, traffic controllers (especially the good ones) will disable the system if two trains are about to be in a conflict. Although the automatic route setting system was meant to enable traffic controllers to look ahead and anticipate, the very same system design made anticipation difficult; this is not only caused by the system's unpredictability, but also because it does not provide any overview of future track use and possible conflicts therein, even though it is the traffic controllers task to solve such conflicts (this will be the topic of the next section). Several sources, such as interviews and an incident report (Inspectie Verkeer en Waterstaat 2009b) report deadlock situations. Apparently, the route setting system does not check whether the route can be continued at a later moment, before it is actually set. In contrast to former practices, the automatic route setting system does not set routes from their desired route endpoints towards the train. This would make sure that a train can continue its route. Instead,

the ARI system just reserves the sections that are mentioned in its plan lines as far as this is possible. In some cases, if a route for another train is set before the first one was finished, it may happen that part of the intended route for the first train is suddenly claimed by the other one. In some cases, this only leads to delays and the situation will more or less solve itself, but in other cases the two trains can meet each other nose to nose at red signals. Although this is safe as long as drivers manage to stop at red signals, it is unlikely that this situation matches what the traffic controller had in mind and the situation can only be solved by driving one of the trains backwards. Although these deadlocks do not occur daily, they do indicate that the system's route setting behaviour is unpredictable.

Despite these automation surprises, there are also advantages that come with the automatic route setting system. Normally, if a train has had to wait it will get its route as soon as it becomes available, whereas in former times a traffic controller regularly forgot that there was a train waiting. Also, drivers report that they are now hardly ever being sent in the wrong direction (e.g. at junctions), which is likely the case because the route setting is inherently linked to the train number.

8.2.5 Automatic route setting creates driver expectancies

Several interviewed drivers complained that the automatic route setting system is in some cases triggered too late. The reason for this is that routes are not claimed very far in advance, so that the route elements can still be used for other trains that might perhaps come earlier. Whether or not an earlier train will show up is actually more a matter of traffic management than of pure chance; and even if there is no conflicting train, in some cases it still results in an unnecessary high number of yellow signal aspects being shown to the driver. 'We have to adapt ourselves to the system, but that should be vice versa', one driver said. In some of these situations, the route setting system is triggered only after the train has passed the signal, but then it usually sets the route, thus giving an unnecessary warning to the driver, who then learns that it suffices to apply the brakes only slightly at this signal. In the minority of cases in which the route really cannot be set, this signal is really indicating the necessity to stop. The driver cannot see that any longer from the yellow signal aspect he just passed – he always passes a yellow aspect – and he may still be braking only slightly, behaviour that is against the formal rules but effective most of the time.

Signal detection theory (which is not about railway signals but about distinguishing information – the signal – from its environment, also called noise) says that providing information in this way will condition the observer to expect a false alarm, thus decreasing the probability of a correct interpretation (Wickens and Hollands 2000). This is something that everyone knows as the fairy tale of the boy who cried 'wolf' too many times. The rarer a signal is, the more information it carries[10].

About half of the local traffic controllers said in interviews that they try to inform drivers when they are going to alter their route; the other half rely on signals to provide that information and deliberately do not inform drivers, arguing that if they did so, drivers would be less alert for signal aspects, which might be dangerous when a traffic controller could not (for whatever reason) inform a driver. Having said this, drivers said that they were almost never informed, and most of them would like to know what the traffic controller's plan is, so that they can anticipate on that. Also, drivers said that they are now always sent via the same

[10] Yet if signals are very rare, they tend not to attract the operator's attention, so that an alarm must be put on them. Such signals may also require specific operator training since the operator can lack skills in such rare situations. However, for displaying yellow signal aspects, this is at the moment not under discussion.

tracks, since the introduction of the automatic route setting system, so that they now never see certain parts of a station which they regard as a deterioration of their route knowledge. In conclusion, drivers say that traffic controllers were more creative before the introduction of ARI.

8.2.6 Setting of partial routes: automatic route setting and accidents

Several interviewees mentioned that accidents appear to occur often in those cases where the traffic controller has overruled or disabled the automatic route setting system. This raises the question of whether the automatic route setting system is somehow able to provide safer routes than the controller. However, from the analysis before, this appears to be unlikely, since the automatic route setting system has the potential to influence not only the traffic controller's but also the driver's expectancies in a negative way, thus making them less aware of what is going on. Another, albeit still hypothetical, explanation can also be given. If there is a conflict between train movements, two trains are competing for the same infrastructure element, and one of them must wait until the other train has cleared its route. The train that will be given its route second in this conflict has a high probability of meeting yellow and red signals. If this train happens to SPAD, a following accident is likely, because the route behind the red signal is indeed occupied by another train. If such a conflict occurs, this means that there is a last-minute decision for the traffic controller as to which train could best go first. As described before, some traffic controllers may then switch off the automatic system so that they remain in control themselves. This does not mean that switching off the system causes the accident. Rather the fact that there is an upcoming conflict is a common cause for both the accident and switching off the system.

Alternatively, the traffic controller could refrain from disabling the system but set a route for one of the trains manually by right-clicking its plan line. The system will then execute that plan line by setting the route associated with it. The intention then is that this train has already claimed its route and thus will go first. But because it is unpredictable when exactly the route setting system will be triggered by the other (intended second) train, it can happen that this train has just triggered the automatic route setting system a split second before the traffic controller clicked the plan line for the first train. In that case, the train that was meant to go second goes first, and the other train will be given a partial route as far as this is still possible. Thus, the manual command to set a route does not result in the whole route being set as was the intention of the traffic controller, but only that part of it that is available. This sometimes results in a route that does not make sense, e.g. starting at one end of the platform and stopping at the other end, or ending at one of the signals that were placed for shunting purposes only (see above). Setting such routes is likely to increase the SPAD probability because it is against driver's (and also the traffic controller's) expectations. This mechanism could be found at the SPAD incidents and accidents in Amsterdam and Sittard (Inspectie Verkeer en Waterstaat 2004b; 2008g). A similar mechanism is to be found if the first route was set successfully (either by the traffic controller or by the automatic route setting system). If the automatic route setting system is then triggered for the second train, it will also set a route as far as possible which can lead to similarly odd endpoints of routes. Examples where this mechanism was present are accidents in Roosendaal, and several accidents in Rotterdam (Inspectie Verkeer en Waterstaat 2005a, 2005b, 2007d). Although the inspectorate has identified that setting such routes is undesired, the mechanism behind it and the fact that the traffic controller is not always in full control, is not mentioned in any of the reports. It must be said, however, that it is a clumsy system design if two operators (ARI and the traffic controller) simultaneously act in the same conflict, without knowing of each other what the other is doing. Although this is just an untested hypothesis, it is well in line with an earlier

report by Railned[11] which indicated a statistically significant increase in SPADs at those stations where the VPT system was rolled out (Griffioen and Vorderegger 2001). Even though the ARI version at that moment could not set partial routes by itself, it could intervene in the traffic controller's manual route setting tasks by claiming a route shortly before, such that the traffic controller unintentionally set a partial route. The hypothesis is also well in line with the fact that the accident rate increased more than the SPAD rate, as was shown in Figure 1.1 and Figure 1.2 in Chapter 1.

In conclusion we can state that:
- the observation that some accidents occur when traffic controllers disable the automatic route setting system is correct;
- disabling the automatic route setting does not *cause* the accident;
- instead, the fact that a route conflict is about to happen, apparently makes traffic controllers disable the automatic route setting system;
- the traffic controller and the automatic route setting system can deal with the same conflict simultaneously, not knowing what the other is doing;
- thus, the automatic route setting system is inadequately designed to support traffic controllers to solve these conflicts;
- the fact that this explanation was not found before indicates that the behaviour of traffic controllers is not monitored and that possible feedback about automation surprises did not result in an improved design;
- the behaviour of the automatic route setting system may lead to senseless routes, even if these were set manually, possibly increasing the SPAD probability;
- these routes are in many cases not set conform former operating practices (i.e. from the far end towards the train) and were in the past in some cases even prohibited.

8.2.7 Discussion on use of partial routes

Setting routes that end soon and thus allow a train to start on a yellow aspect are often said to be set to improve punctuality. Although this is a frequently heard statement, it can be shown not to be true, since it assumes too much intelligence in the route setting system and also it does not match the punctuality measurements:
- The automatic route setting system has no notion of punctuality whatsoever; it is simply programmed to set routes shortly before the train's departure time. In other words, the route is not set to improve punctuality but most of the times basically because it was possible to set it as far as it goes.
- The punctuality argument is often accompanied by a somewhat cynical remark that this is useless, because the train has to stop anyway as soon as it has departed, but it then is said to count for the statistics as a train that departed punctually. This is also not true, firstly because the train would also be counted punctual if it departed two minutes later, so that no partial route is necessary to improve departure punctuality, and moreover because punctuality is only measured on arrival, not on departure. In theory, it is possible that the next train needing the platform can then arrive on time, but at least in those cases where partial route setting led to accidents, no such train was in the vicinity.

Such 'partial routes', in which trains depart on a yellow aspect, make up 10% of all SPADs and their consequences are worse than average (Inspectie Verkeer en Waterstaat 2008h). Therefore, the inspectorate wanted to start a discussion about the relationship between starting-on-yellow SPADs and punctuality (Inspectie Verkeer en Waterstaat 2006). This led

[11] RaillNed was responsible for capacity management and rail safety before its capacity management branch was incorporated by ProRail and its safety branch became the railway inspectorate

ProRail's Capacity Management department to perform an analysis to find out what the capacity loss would be if such routes were not to be set any more (Baas 2007). This report again showed that this subject is cluttered and not well understood:

- First, there can be no capacity reduction when not using partial routes during scheduling by definition. The reason is that when making schedules, the paradigm is that signal aspects should be the best possible and this is not the case if partial routes are used. This means that no partial routes are used during scheduling, and since capacity is normally (but debatably, see section 1.3.4) referred to as the number of trains per hour that can be scheduled, by the railway's own logic no capacity effect should be measurable from not setting partial routes. This also indicates that the system appears to have more capacities in practice than thought by timetable designers, which is now gradually being recognised and made use of.
- Secondly, the report actually did not investigate the effect of partial routes in the sense of starting-on-yellow-routes, although this is what is commonly meant with the term since this is confusing for the driver. Instead, it looked at the effect of splitting plan lines in the VPT system (e.g. a plan line that normally sets a route from track A via B to track C in one go is now split into two separate plan lines, one for each step). However, how many plan lines are used to form a route is totally irrelevant as long as these steps are triggered to be set in time, so that the result is that the driver can see a green signal aspect when he deserves one.
- Thirdly, the question was immediately narrowed down from the relationship between punctuality and capacity, to the question of whether or not it is acceptable to set such routes. A middle course, such as the question how such routes could be set without leading to dangerous situations, was apparently not considered.

In conclusion, the phenomenon of partial route setting appears to be explained from a number of wrong assumptions on how the railway system functions (e.g. technically on the functioning of ARI and on an organisational level how punctuality is measured). Apparently, the level of knowledge and the willingness to correct wrong explanations in the sector is such that these errors are not corrected; instead, these wrong assumptions appear to progress throughout the sector as if they were the real explanation; so that there is apparently insufficient organisational learning.

8.3 Rescheduling infrastructure use

8.3.1 Rescheduling: why and when?

As we have seen in the sections and chapters before, conflicting claims on infrastructure use create difficulties both for drivers and for traffic controllers. Now in order to be able to provide trains with conflict-free routes, it is necessary to have some kind of a plan. This plan can be the timetable, but it can also be the (changed) plan as the traffic controller has it in his mind during operations. In both cases, the 'plan is a number of coherent measures in order to achieve certain goals' (de Leeuw 2001) and 'rescheduling is the process of updating an existing production schedule in response to disruptions or other changes' (Vieira et al. 2003). As such, planning is an essential part of any intelligent control system. It requires a clear goal, and updating the plan becomes appropriate every time an event happens that renders the original plan infeasible or significantly less than optimal, such as an (expected) train delay or an infrastructure failure. In rail traffic operation, planning is essentially a three-dimensional puzzle, the three dimensions being all (parallel) tracks, the trains' positions on those tracks and the time at which they will use each track.

8.3.2 Rescheduling practices

Paradoxical goal
In railways, the original plan is made by the timetable designers, given certain commercial preferences of train operating companies. Any modifications to this plan in response to disruptions are performed by the traffic control department. Since the controllers do not know the original specifications for the plan, they do not have a clear goal, as was indeed indicated from several interviews and literature sources (Lenior 1993; Gouweloos et al. 2003; Regtuijt 2007). Instead, it becomes a pseudo-goal to restore the original timetable as quickly as possible, instead of rescheduling in such a way that the original specifications are met as well as possible. This is ironic since most interviewees indicated that a satisfied customer was their main goal, which was at that moment not even the formal KPI on which ProRail's performance was measured.

Currently, it is the task of the regional traffic controller to reschedule trains network-wide, except in stations, where local traffic controllers do the rescheduling task. This somewhat awkward approach is necessary because each of these actors misses an essential dimension to oversee the whole traffic process: the regional traffic controller cannot see the actually planned track use in stations, whereas the local traffic controller only has a bird's eye view of his own area, but he has no support to oversee how the situation will be in approximately 15 minutes or in neighbouring stations.

'Punctual remains punctual'
Also the rescheduling paradigm 'punctual remains punctual' appears to be an oversimplification which is mainly used to keep oversight and reduce the controller's workload, since this principle does not lead to a minimum amount of delay minutes and it does not make optimal use of the slack that is built into the timetable to cope with delays. 'Punctual remains punctual' then also means 'once delayed will be delayed more and more', since all additional conflicts are then solved at the cost of delaying the already delayed train still more. Apart from the fact that this paradigm is not correct from this more theoretical point of view, it also is not adhered to in practice, since this would mean that hardly any consecutive delays would exist (a consecutive delay is a delay which is propagated from one train to another), but these are quite frequent and make up approximately 55% of all delays (Van den Top 2005; Vorage 2006; Weeda et al. 2006; Weeda and Hofstra 2008).

The current way of visualising the actual situation (Table 8.1 and Figure 2.8) for the rescheduling process is insufficient to create a feasible new plan. What is done now is to keep the original timetable even in the case of train delays and put a remark that train X is n minutes late; for regional traffic controllers, a recent modification is that on their screens the line relating to the train is now automatically shifted in the time-distance diagram. However, this does not result in a new plan that says how this train will be handled on the network and whether any conflicts will be caused by the delay, nor will it be clear whether the delay is expected to increase or decrease given the traffic pattern on the network and the built-in recovery margins.

If the order of trains between stations has to be changed, this is formally the responsibility of the regional traffic controller, although in some control centres practices have evolved where the local traffic controllers manage such situations by themselves as long as the train does not leave the territory of their (combined) areas of responsibility. In standardised cases, so called TADs (Treinafhandelingsdocumenten – Train Treatment Documents) prescribe how to deal with delayed trains and delegate the ability to reschedule trains from regional traffic control to

local traffic control. For example, they indicate that if train *x* is more than *n* minutes late, train *y* goes first. In practice, this does not work because the TAD assumes that all other trains are running to schedule, which they generally are not when the need for rescheduling is most urgent. Since the TADs are also a formal agreement between ProRail and the train operating company, local traffic controllers are assessed on the percentage of times they use TADs correctly. This does not invite them to come up with creative solutions, although creativity is precisely what is asked for if the control measures as defined do not function properly. TADs therefore simultaneously require and forbid creativity.

Regional traffic controllers indicated that they thus focus mostly on large scale disruptions (which are out of the scope of this research) and on rescheduling delayed freight trains; in their opinion, passenger trains are sufficiently flexible so that they will find their way through the network. Also, ProRail reschedules delayed freight trains only in standard paths that were put into the hourly timetable pattern during timetable design. However, during timetabling, these paths are not kept free especially for rescheduling. Although this kind of standardisation has an advantage, since it makes rescheduling easier, it creates problems if this 'one size fits all' approach does not fit on a particular occasion. For example, train operating companies said that they did not inform ProRail if a train's maximum speed was lower than usual, since this would mean that the train could not run at all. Hence it is just put into service without a feasible schedule, since this is the only way to get that train moving.

Not informing drivers about rescheduled trains
Regional traffic controllers, who actually reschedule the trains, cannot contact drivers directly, but only via local traffic controllers or via the control centre of the train operating company. Apart from the fact that reaching a driver is burdensome, also the need to inform drivers is not recognised by most controllers, so that drivers in the end are not informed about changes to their planned times. Thus, drivers behave conform the unchanged plan until they find out, by encountering a yellow signal aspect, that this is no longer possible. They cannot anticipate. Another problem causing this absent information flow occurs in cases of delays, where regional traffic control assigns changed train numbers to freight trains. If the freight train driver does not enter this number into his train radio set (which he generally cannot do, because he is not informed about it), the local traffic controllers cannot contact him.

De Leeuw indicates that planning can function as a straitjacket if it is done at a too high level of detail (de Leeuw 2001). This is also the case in highly used networks. Several interviewees indicated that the timetable as it was designed is changed frequently by the controllers to make it run more smoothly. Many controllers, on the network and on the local level, indicate that the timetable and the infrastructure are too much tailored to the planned situation and leave insufficient room for ad-hoc creativity. Since exact train delays can only be known minutes in advance (Schaafsma 2001; Van den Top 2005; Projectteam Samensporen 2006) it would be wiser not to specify a specific order of trains in the timetable, but instead to structure the track use and time slots such that rescheduling becomes easier during real time operations; of course within the tolerances that are acceptable from a customer viewpoint, which should therefore also be made explicit.

8.3.3 Finding conflicting infrastructure claims

To be able to find conflicting infrastructure claims and solve these by changing plan lines in the VPT system, it is necessary that a traffic controller can have a reliable estimate of the traffic situation in the next (approximately) 15-30 minutes. However, this seems to be not always possible, as was found during interviews, but also during trials for more efficient overtaking of a slow train by a fast train (van Luipen and Meijer 2009). The reasons for the

inability to make accurate predictions will be illustrated below, again using train 2039 (the train used in the example in table 8.1) as an example; see Table 8.2.

Table 8.2: Inaccuracy of train arrival estimates.

Train number	Movement type	Plan time	Delay	Delay measured at (station abbrev.)	Set time	From track	To track
2039	A1	12:15	+ 12	Ed	12:12	AX	23B
2039	A2	12:15	+ 12	Ed	12:13	23B	4
303037	V3	12:10	V+ 10	Ah	12:19	25A	BE
7539	A2	12:21	+ 0	Ahwa	12:19	23B	8

Train 2039 is indicated with a 12 minute delay in Ede-Wageningen (Ed). However, the delay measurement system (based on TNV, the train describer system) gives only on average correct information, but for individual measurements these can deviate up to +/- 2 minutes, depending on train speed and actual track use. This means that the real delay is between +10 and +14 minutes. This delay information is therefore not suitable for process control (Goverde 2003).

A second factor is that it is not indicated to traffic controllers (nor to drivers) how much slack time has been put into the timetable to cope with delays during operations; operators only know how much slack there is by experience. Slack times typically vary between 0 to 3 minutes, dependent on the ability to fit the train into the timetable pattern. This means that the expected arrival delay is between +7 and +14 minutes, so that train 2039 will arrive between 12:22 and 12:29. Because slack time is not made explicit, delays are generally overestimated. Also the VPT system is not helpful in using slack effectively, since its function ‘process delay’ automatically adjusts plan lines to the measured delay, but without taking available slack into account. This means that if a train was measured with a 12 minute delay, it may in some cases be given its departure route also with 12 minutes delay, so that the available slack effectively is not used at all, unless the traffic controller manually changes what the automation changed shortly before.

A third factor that can influence the train delay is that it may lag behind a slower train. In this case, train 7539 is a local train coming from Ede-Wageningen. It is on time to arrive at 12:21, so that train 2039 cannot arrive earlier than 12:23 (given a 2 minute headway). However, if train 7539 were delayed, this would surely affect train 2039. This would not stand out from the information shown in Table 8.2. Note that the delay measurement inaccuracy does not only hold for delayed trains but basically for all trains. It therefore depends on operator experience how well the arrivals can be forecasted and decisions taken accordingly. One driver indicated that if he is delayed, he regularly takes the initiative and calls the traffic controllers ahead to agree on a passing time; by such agreements, he prevents his delay from getting more and more over time; this is however not a very common procedure.

This means that the (expensive) slack time will probably not be used in an optimal way because it is not clear where exactly it is, how much is available and how this could be used effectively: the controllers must spend much mental effort to build a representation of the future situation and derive from the delay information whether there will be conflicts that need their attention. The consequences of possible solutions are also difficult to oversee. However, a change of 1 minute in the expected arrival of a train at a conflict point could or should in some cases already be a trigger to change the order of trains (Van den Top 2005). A two minute change can definitely influence the order since this is on average the amount of time it takes for one train to clear its route. An example where it would have been useful to display conflicts was an accident near Weesp (Inspectie Verkeer en Waterstaat 2008f). In the accident report, the traffic controller declared that there was no need to intervene in the

process because the delays were only a few minutes in size. However, these few minutes meant that a conflict would occur which led to a red signal and, in combination with a SPAD, led to the accident. The conflict, and thus the red signal aspect, could have been prevented easily, since a parallel route was available. It is a formal task requirement for traffic controllers to check the plan for conflicts and to solve these. It is questionable whether this traffic controller was adequately motivated and supported to do so, since the traffic controller could not easily see which of the delays would have a conflict as a consequence. The design of the traffic controller's software support system VPT originally catered for a conflict detection and resolution module, but this was never realised (see section 9.3.4).

8.3.4 Finding solutions for conflicts

Apart from the poor awareness of the expected train delays, interviewees indicated that controllers need to gather information from many sources. The information that is available to local traffic controllers and regional traffic controllers is often static, not always up to date and in some cases inconsistent with information from other sources. An example is the length of a train, which is normally entered into the system during timetable design and as such reflects the planned length but not the actual one. Another example is that a regional traffic controller can look up a station layout, but this is simply a drawing which does not show which tracks are in use. Such a screen is available as well, but it only shows track occupations if a train number is entered, which gives the risk that a track that seems to be free in fact is not. Other information is completely missing, for instance regarding staff circulation plans and whether or not train staff is at the train. Although this is not a matter for traffic control to solve, it does influence their solutions, since it poses restrictions to the possibilities of which trains can move and which cannot. Also the automatic route setting system does not take restrictions into account that a human operator would do or at least would like to do if he could. Restrictions not taken into account are, for instance, planned connections to and from other trains, or staff circulation constraints, or whether a platform track is long enough for the whole train. This also leads to undesired routes being set. The result of all this is that traffic controllers often are reluctant to change the plan because they cannot oversee the consequences of a change.

8.3.5 Management influences

Besides providing the required tools and a clear goal, which we have seen is missing for traffic control, it is also a management task to be clear about responsibilities and to motivate staff.

Responsibility for quality

In former times, there was an important task for NS's central traffic control centre in Utrecht since they were responsible for the quality that was finally delivered. Such a feeling of responsibility now appears to be absent. For instance, if a freight train is delayed due to an error of a traffic controller in station A, the other controllers in stations B and C do not know this and may switch the train onto a side track, arguing that because it is late, it should wait for its next free freight path. Similar cases have been reported for passenger trains, where a train that is once delayed does not receive a priority handling to allow for recovering from the delay, but instead gets delayed more and more. No effort is put into repairing such an error, and thus the traffic control organisation as a whole does not take responsibility for its error.

Another sign that indicates reluctance to take responsibility comes from the interview with one driver, whose task it formerly was to accept or reject the data in ProRail's monitoring system in which traffic controllers register causes for delays. He indicated that very often the category "driver's process" was ticked as reason for the delay, even in cases where it was

totally clear that in fact late route setting was the real cause. Apparently, traffic controllers do not want to be responsible for delays. One traffic controller indicated that he sometimes 'fools' the registration system in order to meet quality criteria. For instance, if a late night service is slightly delayed and passengers would risk missing their connecting train, he would not tick 'wait for connection' (since this would mean that the train had to wait longer than what was formally agreed on with the train operating company, but instead he would then tick 'unexpected high number of passengers' which would still represent a delay, but one for which the traffic controller could not be held responsible.

Motivation
Interviewees indicated that they had quite different ways to reschedule: some stuck strictly to their guidelines and procedures, whilst others interpreted the same rules more flexibly and tried to achieve a little bit extra if a customer wanted it. However, some controllers appeared quite wearied, because they had too often tried to propose solutions which were then not picked up: 'Formerly, I used to think about improvements and proposed them, but I stopped doing that'. This weary impression reflects the majority of the interviewees, although it appeared to be dependent on the control centre where a respondent was working: some appeared to be very motivated to keep on trying to improve, but others appeared paralysed and only did what was formally required. The discouraged ones apparently gave up suggesting improvements and now just stick to their formal task description. Whether they are still motivated or discouraged, in both cases this indicates that these people do in fact want to create high quality services for their customers. That they gave up doing so in some cases, is not a very healthy sign and definitely needs to be addressed.

8.3.6 Summary of causes for ineffectiveness of rescheduling

The high amount of consecutive delays indicates that either the plan is not correctly rescheduled by traffic controllers, and/or it is not correctly executed by the automatic route setting system and drivers. This can have several reasons. As far as incorrect rescheduling is involved, these are:
- Not all conflicts are looked for and solved, e.g. those of passenger trains;
- Not all conflicts are found, e.g. because they are not explicitly indicated;
- There are insufficient possibilities to react which results in conflicts not being solved. Limitations can be either physical in timetable and infrastructure design or it requires too much mental effort to oversee the consequences and implement changes.

Another reason is incorrect execution of a rescheduled plan:
- The automatic route setting system is not programmed to execute 'on time remains on time' but to execute plan lines such that the planned order on tracks remains the same, which is not in line with the regional traffic controller's practices.
- Drivers are not informed about the changed plan and act according to their original schedule.
- The local traffic controller decides to deviate from a rescheduled plan that was made by the regional controller.

The question remains what the added value is of a controller, if he can neither completely oversee the situation nor the consequences of his own responses, and therefore sticks to the original timetable and the paradigm 'on time remains on time', whereas it is basically his task to change the timetable to cope with disturbances? This is not an assault on the controllers, who without exception are, or at least were, very motivated to do their work. It is rather a question to ProRail's management, since it appears that many of these controllers could perform much better than they now actually do. This requires a regular analysis of the local

traffic controller's and regional traffic controller's practices; not in the first place to see whether they stick to the rules but to check whether the procedures as they were thought out do indeed work in practice; and whether they do not have unintended side effects. This asks for learning at an organisational level.

8.4 Traffic control human factors

8.4.1 Automation versus human operators

The previous sections showed several examples which make it difficult for the traffic controller to predict the future situation. As Roth et.al. (1999) indicate, dealing with unplanned demand and dynamic replanning is a cognitively difficult task. Successful performance depends on the ability of traffic controllers to monitor train movement beyond their territory, anticipate delays, balance multiple demand on track use, and make rapid decisions. Support by technology and automation may ease this task, if it is applied correctly. This section will indicate some traps and indicates possible directions for improvements.

Several authors have addressed how automation and human operators should and should not be combined. Lisanne Bainbridge (Bainbridge 1987) wrote an article about the ironies of automation. She writes that 'the classic aim of automation is to replace human manual control, planning and problem solving by automatic devices and computers'. Often, the operator then gets a monitoring task, which, however, is a task which humans cannot perform very well (see section 7.4). Monitoring an automated process and taking over if necessary is associated with several problems, such as:

- Inability to monitor the computer's performance if the computer can think and act faster than the operator can (which is often a main argument to automate the process);
- Loss of cognitive skills, e.g. understanding how the process behaves, knowing that situation X may result in problems, or knowing what to do if situation X occurs. Thus, the operator gets less skilled in understanding when he should take over;
- Loss of practical skills. Physical skills deteriorate when they are not used, for instance fine-tuning and timing of actions get worse. Thus, the operator is less skilled once he has taken over from automation;
- Problems to take over manual control since the actual system state and its dynamics are not yet 'loaded' into the operator's mind;
- The human being must know which tasks the computer is dealing with and how. Otherwise the same problems arise as in human teams in which there is no clear allocation of responsibility. Since the operator must not only have a mental model of what the process under control is doing, but also how the automation will interfere in that process, a higher rather than lower workload may result from automation.

These findings are not new. Before the VPT system was introduced, one of the engineers of the Dutch railways proposed the following assistance to Dutch traffic controllers (Baud 1984):

- the normal process can easily be automated, especially gathering of standard information;
- during a lightly disturbed process the system should not work automatically, but only propose ideas to the traffic controller (basically to assist the traffic controller and to avoid that he would mistakenly (due to loss of oversight) choose an option that makes things worse;
- in a heavily disturbed process, the system should only provide information about the current situation (e.g. display actual train compositions and positions, set passenger

information displays only at the last minute so as to be sure that no incorrect information is shown earlier).

The problem with such ‘assistance’ is that, by taking away the easy parts of his task, automation can make the difficult parts of the human operator's task more difficult. Research was done to see how tasks could best be distributed between men and computers, to avoid this phenomenon that was later called ‘out-of-the-loop’ performance problem. It turned out that operators were slower to perform tasks manually if they had to take over from automation, compared to the situation where they had been working manually before. Also, workload did not really decrease when automation was introduced, but the comprehension of detected information (level 2 situation awareness) decreased (Endsley and Kiris 1995). This even led the researchers to the question whether automation of cognitive tasks should take place at all. Broadly speaking, it was advised to keep the operator in the loop by letting him make decisions. Implementation of such decisions could then be executed by automation; automation could also help for producing suggestions on how to solve problems. Later research indeed confirmed that operator situation awareness would improve if options for improving the situations were proposed by automation so that the operator did not need to spend mental effort on generating strategies. The same research indicated that another beneficial option is if the operator can queue tasks in advance and automation executes them whenever appropriate (the way in which the automatic route setting system is currently operated). Although this works fine as long as the automation does its work, a drawback is that it takes longer to recover from automation failures: the operator’s mindset is focussed on the future, so that a mental representation of the present state must then first be built up before the automation failure can be handled (Kaber and Endsley 1997).

Although no study is known in which the Dutch automatic route setting system was systematically analysed with regard to this respect, the British ARS appeared to have some of the properties of ‘clumsy’ automation, yet it provided the lowest workload, so that the advantages of automatic route setting should not be forgotten in the light of its weaknesses (Balfe et al. 2009).

8.4.2 Requirements for information displays

Endsley (1999), the brain behind the concept of situation awareness, provides some requirements for representation of information in order to enhance the operator’s situation awareness:

- Information should be provided such that the operator does not need to combine data from several sources or make calculations in order to understand what data presented actually means;
- The data should be coherent to the goal(s) the operator must achieve;
- In the case of critical events, a cue should be provided in a salient manner;
- The information should provide oversight which is broader than that of the operator’s goal alone to provide the operator with some general understanding of the system he is working in;
- There should be a support system which provides the expected future state of the system.

When we compare these requirements to the practices described in sections 8.2 and 8.3, we see that on all of them improvements are possible and necessary. Other suggestions for improvement are:

- Let operator’s work manually for a short time each shift, to maintain their skills; if this is infeasible, use simulator training (Bainbridge 1987).

– In order to provide more time to have timely responses from operators, this requires some form of predictability. Therefore, designers must make sure that people can anticipate the automata's behaviour (Woods and Hollnagel 2006). Repeated human errors can therefore be considered to be a consequence of a poor system design (Leveson 2004). This appears to be the case also in the design of the automatic route setting system, which is unpredictable for its user.

An example of a traffic control support system in which most of the abovementioned recommendations were taken into account was recently designed by the Uppsala University[12] (Kauppi 2006; Isaksson-Lutteman et al. 2009). Their design is a screen in which nearly all relevant data is combined: planning and realisation of the rail traffic are shown by means of a time-distance graph. Also conflicts in the plan, maintenance timeslots and actual information on train properties are displayed, so that the operator has a good oversight of the situation now and in the short term future. The automatic route setting module does its work only if it fits the current plan of the controller. Since the problems mentioned in their research resemble the Dutch ones nearly completely, it is suggested that a new traffic control system adopts at least the same approach, but we will see in Chapter 10 that even more improvements may be necessary.

8.5 Ambiguous practices: cooperation with other actors

8.5.1 Five paradoxes

We now turn from problems of interaction between the controllers and their technology interface to problems associated with their person to person interface. Traffic control, as a task to (re)schedule means of production in the short term, inherently requires communication to see whether all requirements of a solution are met but also to inform others in the control process about disruptions. ProRail's local traffic controllers not only have to interact with regional traffic controllers in their own organisation, but also with drivers and rescheduling staff from train operating companies. Although there is a formal interaction model (the 'control square') this is an oversimplification as Figure 8.3 shows: most controllers have to interact with many more actors than the control square suggests.

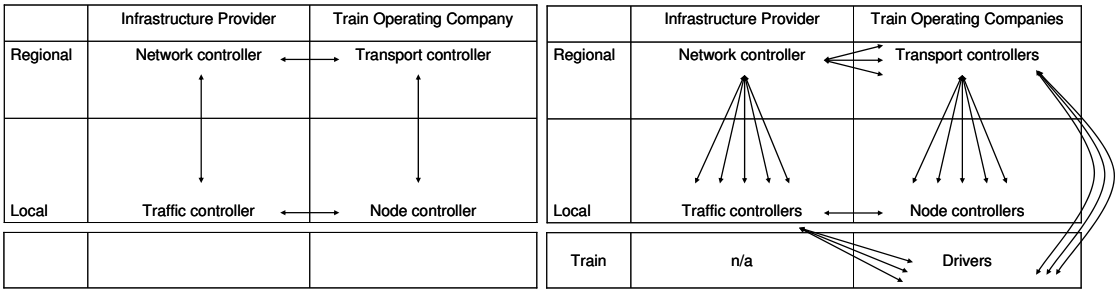


Figure 8.3: Control squares as they are defined formally (L) and used in practice (R).

This section is a shortened and somewhat updated version of an article that I wrote together with Bauke Steenhuisen, who performed research within ProRail independently of my own research and came to similar conclusions (Van den Top and Steenhuisen 2009). Both research projects led to several paradoxes related to communication being identified:

1. Operators occasionally lose (their sense of) control over the situation. They then retreat to 'islands' in which their job becomes tractable again, for example by limiting themselves

[12] More information, amongst which working demo files can be found on the following website: http://www.it.uu.se/research/project/ftts

to their formally defined tasks and not trying to be proactive, problem solving and creative. Because all tasks are tightly interconnected, a fragmented optimisation process with often suboptimal system output may result.
2. Communication creates constant painful trade-offs. On the one hand, interaction between operators is necessary to get information vital to a decision and, more basically, to make sense of the situations in which operators find themselves. On the other hand, operators do not want to interact when communication is very time consuming and does not pay off, for example because it increases uncertainty. The effect is that operators tend to present obstacles to each other instead of engaging in a mutual optimisation process.
3. As operators sit on their islands, they all develop different workarounds for similar problems. Some bend the solution to fit the rules, some bend the rules to fit the solution.
4. Retreating to islands is caused by both a lack of information and too much information. In the latter case, information is not relevant, not up to date or not specific for the activity of the operator.
5. Too many managerial interventions and too little managerial support make operators more inclined to retreat to do only what they are formally asked to do, thus forming these islands and not optimising the ensemble.

In the following sub-sections, these paradoxes will be illustrated. They need to be at least recognised and understood; in some cases, the paradox can then be solved, in other cases this appears to be impossible, yet one can try to shift the negative consequences to the place where the harm is the least.

8.5.2 Paradox 1: A cascade and an archipelago

From interviews it could be concluded that the centralised approach described as the cascade model in Chapter 2 partly exists in practice but it is also partly fictitious, since co-operating professionals, in fact, form not only a cascade but also an archipelago. Planners and operators each sit on their own island and push and pull the operational processes from their own stance. The way different tasks cooperate is not organised. On paper, the cooperation has an organisational model, but many operators retreat to their islands in practice as they feel that things do not work as they should. They then use methods that are not recognised by the management or even prohibited. Particularly people from the time that the railways were one company know these unofficial solutions, for example, last-minute adjustments to the planning. These solutions can help a traffic controller escape from straitened circumstances by making possible what is formally impossible.

In some cases, traffic controllers find themselves in a situation in which the planned track use for certain trains is such that this will easily lead to conflicts when small delays occur, and where an alternative route could have just as well been chosen. Some traffic controllers change the plan and make things work. Others, who have lost their motivation, just leave the plan as it is, arguing ’if they plan it this way, they will get it this way’.

As was already shown in Chapter 7, train drivers work in isolation cut off from necessary information that could in some form be made available from the traffic controller. Also at a regional traffic control centre, we observed many examples in which regional traffic controllers work without fully considering their colleagues. For example, a traffic controller assigns an unrealistically fast train path ad hoc to a freight train. This train path is the only one available in the short term. Two regions further on, the freight train comes into conflict with other trains. Checking the feasibility of such ad hoc plans with local traffic controllers is a formal requirement but in practice it is sometimes tacitly assumed that there will be no objections. Traffic controllers are often confronted with an overload of information and

communication. They do not always ask all concerned colleagues before executing a plan. The burdensome communication is avoided wherever possible, sometimes at the cost of a suboptimal plan or creating a local impossibility for someone else to solve. There are hundreds of traffic controllers and transport controllers simultaneously at work respectively for ProRail and the rail companies. They all risk developing plans in parallel, which might turn out to conflict with each other in practice. For instance, ProRail's traffic controller may want to cancel a highly delayed train service, since otherwise other trains would be affected too much. However, the train operating company wants the delayed train to be squeezed in somehow, accepting that its other trains are delayed as well. The reasons to accept these delays can be invisible to the traffic controllers of ProRail. The train can be crucial in offering enough seating capacity or in managing the planning of train staff circulation. As a final example, a train company may strategically withhold information on an extra freight train that needs to cross the whole country. As the train suddenly appears at the border, the traffic controller tries to solve the matter and the train company is helped. So, the cascade and the archipelago both apply when describing the actual practices.

8.5.3 Paradox 2: Operators want and do not want to interact

The island-mechanism plays a role in how operators understand what they need to do. The island-mechanism is also a matter of seeking legitimacy in a practice that is not fully understood. A local traffic controller describes how he responds to the decisions of others:
'A regional traffic controller often makes dubious decisions. We do not agree, but we follow the instruction. It is not our task to come back on decisions. In turn, we do not see the consequences they see. So, we let it pass.'

When discussing the complications that operators daily face, they often explain: 'nobody tells me anything'. Operators feel themselves forced to act unguided in an uncertain environment. In response, operators take decisions taking account only of those actors whom they really need, whilst blocking the others – who then can also say 'nobody tells me anything'. Some operators are also forced onto their own island if their colleagues do not interact with them. Observations in an emergency call centre showed that directing another's attention to something depended on the ability of an operator to see whether his colleague was interruptible (Woods and Hollnagel 2006). Such clues have been disabled in the current traffic control organisation, where different actors are now located in different control rooms: they can no longer see each other, thus requiring them to communicate by phone, which may interrupt the colleague-controllers in their normal work. Also, it has made decision making much slower, as was pointed out by one traffic controller: 'What took 7 seconds when we sat in one room, now takes 7 minutes'. This situation will be improved, however, when ProRail's new traffic control structure comes into service (see Appendix A4).

Also, traffic controllers cannot see whether the way in which they set routes leads to a situation which requires drivers to slow down, in some cases unnecessarily. When drivers contacted traffic controllers on occasion to compliment them on the creative way in which they had solved a conflict, several cases were reported where traffic controllers assumed at first that the drivers were kidding. This indicates the sphere of guarded mistrust present in such cases. For reasons of clarity, operators have a strong interest in the advantages of working apart and a great aversion to the disadvantages of working together. Communication in prime time makes workers pay a high price for something that is not directly valuable. As a result routines arise to work without feedback. This explains the archipelago, and so, the interactions among operators are not directly driven by clear incentives or a clear model. The emphasis on working apart appears not to be optimal but it is at least understandable. How operators make decisions deprived of feedback is the subject of the next paradox.

8.5.4 Paradox 3: Operators decide by clear rules or not

To understand how operators make decisions we need to look beyond the rules and top-down incentives they receive. Brehm and Gates (1997) use the distinction between 'central' and 'peripheral' (communication) routes to persuade a person to act. Top-down planning and rules in the communication between operators, for example, are central routes. Peripheral routes are unofficial routes in which operators recognise a cue and take a shortcut towards a response, consistent with their past responses or responses of their colleagues.

In this study we met a cocktail of examples with a whole range of both central and peripheral routes. This cocktail makes the way operators behave quite unpredictable for their colleagues, thus again encouraging them to operate on an island. An example of a peripheral route is a traffic controller who has maintenance crew working on one of the tracks. It happens that he knows the crew member who is responsible for the crew's safety from earlier days and decides to keep an eye on him, as he knows that that person is not too careful with some safety procedures.

Another example where decisions are made following clear rules (or not) is that of safety and reliability. The clear rule is that safety comes first. The fuzziness starts as soon as there are different opinions about what is safe. Some traffic controllers argue that the safety system assures that the correct signal aspects are being shown and that it is the task of the train driver to obey these, and they set routes in such a way that trains can run as punctually as possible. They are, thus, following the central route, retreating to their island and sticking to the formal rules, in which they do not have to inform the drivers. Other traffic controllers argue that their own actions have consequences for the drivers or maintenance crew in the field. They would not set certain routes or do so only after they have contacted the driver, if they think that this is necessary to improve safety or punctuality. Although this is not prescribed, these traffic controllers have developed a heuristic and a more pro-active attitude.

8.5.5 Paradox 4: Information overload and scarcity

In order to make the right decisions, all actors involved need information that is valid and relevant to their task. Much information is available, but not always at the right place. For instance, a traffic controller has to contact a driver to ask for his train length or the load, if he needs to know. This creates either uncertainty or time-consuming communication procedures. Also, it disturbs the driver in his safety-critical task while the information must be known also at some other point in the system. Currently, communication effort has to be spent in finding out what is actually going on, instead of using that effort to take decisions about how the situation should be changed. To save communication and to reduce complexity, practitioners develop informal routines that need as little communication as possible. But this burdensome communication is, at the same time, very rich and necessary to optimise the many value-conflicts that are spread over many operators.

For example, a planner makes some changes in the computerised planning, because one track on a double track line has to be closed for maintenance. Although the changes are also available to the traffic controller, he will not know these in detail unless he looks them up explicitly. The traffic controller may thus make some adjustments of his own, with the standard plan in his mind instead of the changed plan, and is suddenly surprised that two trains come to stand in each other's way. Another example: a regional traffic controller changed the timetable of one train, in order to solve a conflict between two train movements. This information did not reach the driver, because the regional traffic controller cannot contact drivers. Thus, the drivers were executing the old plan and still arrived at a station at the originally planned time and had to wait until they could proceed. We see that there is

much information available at different places in the system, but information flows are not always conceived from the viewpoint of the user of the information. Since it is a task of management to agree and decide on system and procedures that are used to communicate, it is interesting to see, in the next paradox, how managers facilitate or interfere in operational processes.

8.5.6 Paradox 5: Too much and too little relevant management

Managers feel that it is hard to make a difference for operators facing uncertainty. When operators cannot assess the situation they are in, neither can managers. In the same situation, managers compliment operators who pragmatically deviate from rules as well as operators who strictly comply with the rules. Rules are one way in which managers try to get a grip on the operational processes. However, as we have shown, the large amount of rules makes it difficult to keep oversight on how the system behaves. More rules make operators stick to their own operating procedures and optimise their own task rather than the ensemble, at least if these are rules that prescribe what to do and what not to do, instead of rules that indicate what the operator should achieve.

Many operators complained that managers insufficiently address the ideas they put forward, yet managers explain that they are busy with a multitude of projects including many ideas of traffic controllers. As operators receive little direct feedback, however, it seems to them as if their ideas are not picked up. Again, this is an incentive for operators to retreat to their own islands and do only what they are asked to do. It makes them refrain from suggesting improvements, which, again, indicates that the railways need to empower their organisational learning.

Given the importance of informal methods and communications it appears to be useful to extend knowledge about traffic controllers' behaviour in more detail. A British study recorded regional traffic controllers on video tape, which revealed many interesting insights into how they perform their tasks and which informal coping strategies the operators use to overcome apparent problems. By further investigating operator behaviour, the design of the control room and the operator support systems could be further enhanced to support this behaviour, ease the problems and enhance overall performance (Farrington-Darby and Wilson 2009).

Another option for studying these phenomena is to use Cognitive Work Analysis (Vicente 1999). Whilst task analysis methods focus on normative behaviour, i.e. on what the operator must do, work domain analysis also takes into account the effort that operators spend to cope with unplanned events which take place in complex socio-technical systems. An application to rail traffic control has been performed by Millen et.al. (2009).

8.6 Traffic control in safe envelope terms

8.6.1 Overview of information provided

As in the previous chapter, we will link the findings in this chapter to the safe envelope requirements mentioned in section 4.8. The traffic controller's task is to modify the operational and viable envelopes of each of the trains under his control. Again, we will see that the operators are missing essential information to perform their tasks.

The safe envelope concept was derived to see how the energy in the system is controlled and how the system's goal can be achieved. Traffic controllers can influence the energy carried by the trains only indirectly, via train drivers, but they can have a role to provide information to

drivers and to make sure that the train's operational envelope is updated whenever required and possible. This means that traffic controllers should be able to see information about the basic information elements for each of the trains in the area for which they are responsible. In section 4.8 a number of such information elements were derived from the safe envelope concept. Table 8.3 provides an overview of the very limited extent to which current local and regional traffic controller's task support fulfils these requirements. As in Chapter 7, the table is split into traffic management (the planning and rescheduling task) and infrastructure allocation. Allocated infrastructure is the outcome of the rescheduling process but it can also pose limitations to rescheduling if certain infrastructure elements remain unavailable at a moment that they were planned to be used again.

Table 8.3: Overview of safe envelope requirements for traffic management.

	Rescheduling (time/spatial domain)		Route setting (spatial/physical domain)	
Information item provided	Local traffic controller	Regional traffic controller	Local traffic controller	Regional traffic controller
Actual state	No (no explicit insight into the time dimension)	No; only implicitly (plan + deviations), no exact track use or conflicts in the plan	Yes (routes set + train locations)	No; no information on track use.
Goal state	No (plan becomes goal)	No (plan becomes goal)	Yes; oversight of trains for which route setting did not take place in time	N/A (route setting is not a goal but a required precondition in order to reach the goal)
Operational envelope	No, the plan contains no explicit tolerances	No, the plan contains no explicit tolerances	No; cannot see which alternative routes are also acceptable.	No
Attainability envelope	No, the plan contains no explicit information on buffers and margins	No, the plan contains no explicit information on buffers and margins	No, there is no updated real time feasible schedule	No, there is no updated real time feasible schedule
Controllability envelope	No information provided on last moment to solve a conflict (e.g. before train reaches distant signal)	No information provided on last moment to solve a conflict (e.g. before train reaches distant signal)	No information provided on last moment to set a route (e.g. before train reaches distant signal)	No information provided on last moment to set a route (e.g. before train reaches distant signal)
Viable envelope	No, but only required in emergencies; otherwise n/a	N/A (responsibility of traffic controller)	No, but only required in emergencies; otherwise n/a	N/A (responsibility of traffic controller)

8.6.2 Traffic management and infrastructure allocation as a DDA control loop

It is salient that most information elements are not or only implicitly available to the controllers of the system. This corresponds with the earlier notion that these controllers appear to have only limited situation awareness, and as Conant and Ashby (1970) pointed out, process control can, at its best, only be as good as the control model which controllers have to represent the system and its behaviour. In the case of traffic control this has the following consequences.

The ability to detect the actual situation is divided between local and regional traffic controllers: although both have a task to (re)schedule the track use some time ahead, which requires insight into the three-dimensional time/distance/track-domain, one of them can only implicitly oversee the time dimension, whereas the other cannot see the track use in detail. Either this requires much communication and leads to long reaction times, which means that certain disruptions cannot be responded to quickly enough, or it means that some information is not taken into consideration when a decision must be taken quickly, thus risking a suboptimal solution or a problem that will arise later at a different location.

Diagnosing the situation is difficult because delay measurements are imprecise, and also the behaviour of the driver and the automation route setting system is difficult to predict because it is not coordinated. This leads to delay estimate errors between -2 to +5 minutes; apart from that, relevant information is not the absolute train delay, although this is easy to measure, but whether such a delay will lead to a problem, e.g. breaking a passenger connection or leading to an infrastructure allocation conflict. This information is currently not provided and hence, little can be done to avoid such problems unless the (team of) operator(s) is both experienced enough to foresee such a problem correctly from the limited information that is available and is also intrinsically motivated to avoid it. Presenting information in a way that is more relevant to the task the operator should fulfil, will help the operator to understand what he must do and will also reduce mental workload to understand the current situation, thus freeing up mental capacity to think of possible solutions. An example from a Swedish study is a screen where conflicts are marked with yellow squares and circles (each indicating a different kind of conflict) and the task of the traffic controller then becomes to eliminate all yellow squares and circles, instead of having first to find out where these marks could possibly be (Kauppi 2006).

The result of the current lack of information and situation awareness among controllers is that the process cannot be controlled on a detailed level more precisely than in whole minutes, and it is slow to react to sudden disturbances. Also, even if a solution is found, it takes relatively long to implement it: it is impossible to set a route directly but it must be programmed into the VPT system first. This was intended to have the advantage that other operators could see from the same information system what is happening and several information systems would be coupled to the VPT information source. Nevertheless, the solution would be implemented quicker if a traffic controller could immediately send a train from track x to track y and the information in the VPT system would then be updated accordingly once this action was performed.

8.7 Conclusion

This chapter has dealt with the question how the traffic control process influences quality of service and safety, or, in other words, how it succeeds in supplying conflict-free train paths. Several causes could be identified that indicated that this process is not optimally supported.

The effort of a traffic controller should not be spent on finding out what the situation actually is and where problems are about to occur, but into selecting solutions to solve these problems in advance. Currently, operator experience fills the gap of incomplete information, but this leads to different and in some cases incompatible working methods and hence to an unpredictable quality of service: for example requests from train operating companies that are not considered as a problem for one controller, may be considered to be one for someone else. Every actor is dependent on someone else before he can make a decision. Agreeing on a strategy can take more time then is available to let the traffic process run smoothly, so that the control process lags behind the real system and the system becomes unstable. Also, the behaviour of the automatic route setting system leads to confusion, since two controllers – a human and an automated one – can work on the same process without consensus about what to do first.

Several paradoxes were described, which can, however be considered rational and understandable responses to the current work circumstances, at least to a certain extent. A paradox generally indicates that there is insufficient understanding of the underlying logic, because a paradox is a situation with a result that one would not expect on the basis of the current mental model, but which can be understood if the reason for the unexpected outcome is understood. By showing both sides of the paradoxes, this chapter intends to provide enough information to understand these paradoxes and solve them incrementally by making their underlying mechanisms work better. Also, it is necessary to monitor whether renewed processes do lead to their intended effects, do not lead to harmful side-effects and whether they can be made more smoothly. This requires that experiences and behaviours from operation are fed back to designers and management, both before and after the change.

Without denying the usefulness of the expertise that network and traffic controllers have, and which is important to have, its *necessity* indicates that control of the traffic process is insufficiently supported: if it is impossible to see a problem or to know how to solve it without expertise, this implicitly indicates that support systems fail to indicate which problems need attention and what the consequences of possible solutions are. If the operator is to solve conflicting infrastructure claims and to do so more than 15 minutes in advance, why does he not have tools that provide him with an oversight of the situation as it is expected to be in 15 – 60 minutes so that conflicts can easily be found? These are currently not available, which is leading to poor situation awareness of both local and regional traffic controllers.

Apart from that, not all conflicts will be found because regional traffic controllers tend to focus on freight trains only, whereas local traffic controllers are supposed not to change the plan in the last 15 minutes. The result is that conflicts with passenger trains and conflicts that arise in the last 15 minutes before operation are relatively likely not to be solved. Also execution of the plan may lead to surprises, since in some cases the automatic route setting system behaves unpredictably so that routes are set in different ways than the traffic controller actually intended. When this leads to routes that are illogical for drivers this is likely to increase the SPAD rate, for instance when routes are unnecessarily set leading to shunting signals or from one end of the platform to the other. Also, drivers may not behave according to the changed situation, since they are generally not informed about such changes. The signalling system, and especially the yellow aspect in it, does not provide such information, as we saw in Chapter 7.

In summary, the following causes for quality and safety problems were identified:
- Information necessary to control the process can be outdated or missing: there is no (precise) information available to controllers about the actual speed, position and length of trains;
- Insight is available in only two of three essential dimensions for rescheduling. Local controllers only have a spatial view, but they miss a clear insight into the time dimension. Regional controllers do have information on the time dimension, but not into the exact track use;
- Information must be gathered from several sources;
- Time consuming communication with other relevant actors is necessary;
- There is complexity due to unforeseen decisions of other actors;
- Complexity is introduced due to unpredictable behaviour of the automatic route setting system;
- It is difficult to make a precise estimate of the system's future behaviour;
- There is no indication of tolerances within which the process should be maintained;
- There is a focus on setting routes, not on solving conflicts.

Nearly all of these causes were also found in a Swedish study (Kauppi 2006). This gives rise to the belief that the problems mentioned are rather general to railway operation and may also exist in other countries. In that case also solutions from different countries could be exchanged, although it should be made sure that not just the solution is 'bought' but that there is primarily sufficient understanding of which problems must be solved and to what extent such solutions help in doing so; some modifications or extensions may be required.

In order to control safety and quality of service, it is necessary that the traffic controller and the drivers in his region share the same plan, so that trains can be kept away from yellow and red signals as long as these are not cleared. This would make the traffic process more predictable both for drivers and traffic controllers; also, the signalling system then can be used for its originally intended function, since it can then serve to indicate that there is a planning error requiring an unplanned stop. Developing such solutions was and is a task of system designers and management: how they do this will be discussed in the next chapter.

9 Influences from design

9.1 Introduction

The previous two chapters showed how operational process control works at the moment and which problems arise. It was shown that drivers as well as traffic controllers have only limited support in their tasks: goals are not always clear, neither are the boundaries within which they can or should operate. Their attention is not always focussed on what requires attention, such as the adherence to the planned train path or red signals in the case of drivers, or the upcoming conflicts in infrastructure use in the case of traffic controllers. Thus, these actors are not well supported to perform their DDA loops, which leads to a limited situational awareness. Many of these problems have their roots in earlier choices, such as the tools provided (or changes therein) and the working procedures that are prescribed.

This chapter will therefore describe how choices on the design level are made and how this influences and supports operational processes and their results. The chapter consists of a large number of cases. Each case is described in a separate section of the next chapter, mostly in a standardised format consisting of three parts. The first part describes the intention of whatever the topic of case is, then the practical implications are discussed and finally, lessons are learned in this thesis.

The cases are ordered in three sections, of which the first (9.2) contains cases concerning the external function of the railways: the service offer to their customers and the role they play in society. Then, cases concerning the design and support of operational processes are dealt with in section 9.3. Section 9.4 deals with the design of safety systems, which are required in those cases where the operational processes need some backup in order to prevent accidents.

In fact, these cases serve as some feedback in the cascade model. Please note that the main topic of this thesis is real time control, so that the analysis below neither claims to be complete with regard to design influences, nor does it go in-depth to the extent necessary to understand and resolve all problems at these levels. As was described in section 6.3, when interview results or other sources used for the theory-in-progress pointed out that certain topics at the design level appeared to be relevant to (partially) explain the phenomena faced by drivers and traffic controllers, these topics were added and additional interviews or

document analyses were performed to elucidate them. These underpin the cases described here, which all fit the following selection criteria:

- To be able to assess the effectiveness of new or modified designs, the cases should cover a broad variety of design influences. Attempts were made to find cases for each of the functions in the cascade model. The high number of examples necessarily limits the depth of analysis, but preference was given to identifying whether design influences give a consistent picture throughout the whole cascade, rather than giving a highly detailed validation for one or two cases.
- The cases should be relevant for rail safety and quality of service during real time operations;
- The researcher should be able to get sufficient information about the case;
- It must be possible to find causal relations to confirm or invalidate the theory-in-development;
- The case should come from applied designs, since otherwise no knowledge from practical experiences can be gained (additional cases that are not yet fully applied are described in Appendix A, in order not to lose the partial insights they also offer).

After the description of the cases, twelve recurring underlying mechanisms for quality and safety problems arising from them are summarised as the final stage of the grounded theory approach. This also provides 'handles' for management to improve the situation. These, then also serve as a basis for improvement, the topic of Chapter 10.

9.2 Designing the service offer

The cases in this section relate to the relationship between the railway system and the society. What do the railways offer and how does this contribute to high quality services?

9.2.1 Conservative philosophy towards innovation for utilisation

Intention
As was introduced in Chapter 1, the railways want to increase their network utilisation by means of innovative processes. This means that functions throughout the cascade model may need to be changed and their cooperation be improved.

Practical use
With regard to implementation of utilisation measures, the following pinch points (amongst others) for successful implementation were identified by Eindhoven Technical University (Fischer 2004):

- Sub optimisation, because all actors in the industry focus on their own processes rather than on the ensemble;
- Conservative attitude, which inhibits innovative ideas. The railway industry sticks to standards for which the reason has often been forgotten;
- Lack of vision: actors in the railway industry tend to perform mostly short term activities, which cannot solve the aforementioned structural problems the industry is faced with.

ProRail's former innovation manager Dijkstra (2007) indicated that the amount of passenger kilometres produced per invested euro in the railway system has been constantly decreasing since the 1960s, or, in other words, that the railway system has become more and more expensive to achieve the same result. Indeed, whilst the inflation between 2000 and 2009 was 24%, the fare prices increased 51% (Trouw 4 aug 2008).

Lessons learned
Obviously, there is a need to innovate in the railway sector, but as Dijkstra indicates, there are few incentives for the railway sector to innovate, which is especially true for operational departments. He says that this can be explained by the fact that making mistakes in railway operations is clearly not desired, whilst innovations require a certain risk to be taken. In the next example, we will see how this works out in practice in timetable design.

9.2.2 Public timetable design: reliability comes at the cost of longer travel times

Intention
An attractive service offer is a prerequisite for the functioning of any railway system. A total redesign of the 2007 timetable led to much public debate. Why was this redesign necessary and how well did it succeed in providing an improved timetable?

Due to additional trains and opening of new stations, the 1970 timetable structure (see section 2.3.5) came under stress. The required running times between hub stations could not always be met in practice, which led to poor punctuality and scheduled connections which had to be broken. The 'Benutten en Bouwen' project (Project Benutten & Bouwen 2003a) proposed a metro-like transport network, on which trains would run with high frequencies and on fixed corridors. In 2004 a team of ProRail and train operating companies started to redesign the timetable from scratch. Another reason for a total makeover was the planned opening of three new lines in December 2006, which would require significant modifications to the timetable anyway. These were the so-called mega projects: the high speed line between Amsterdam and Brussels, a quadrupled track section between Amsterdam and Utrecht and a dedicated freight route between the Rotterdam harbour and Germany (BetuweRoute).

Practical use
In the design of this new timetable for 2007, which came into service in December 2006, the travel times between many nodes were made longer in order to assure a higher train punctuality. However, this meant that many trains were no longer synchronised in stations and travel times have therefore increased. This was shown by Peters (2006), who retrieved the quickest travel times between 71 stations automatically from the NS website and compared the travel times during 2006 with the new 2007 timetable. He did so both for peak periods (departure between 07:00-08:00) and off-peak periods (departure between 13:00-14:00). This

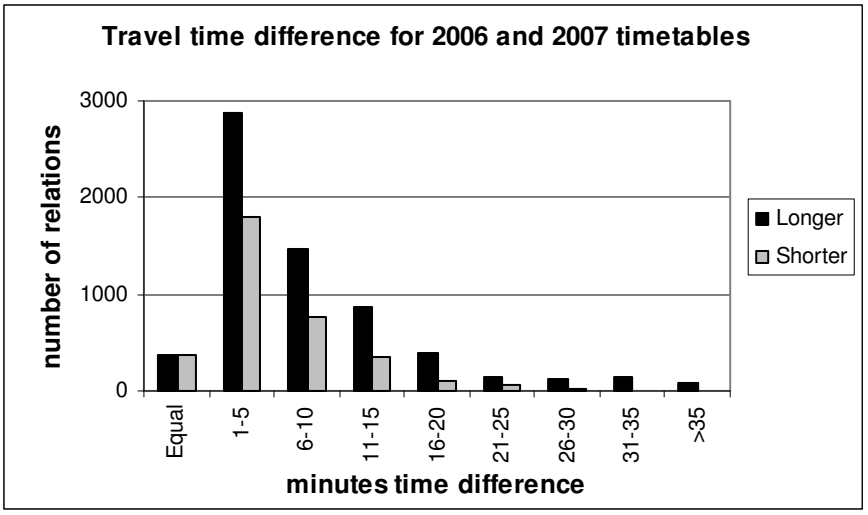


Figure 9.1: Travel time differences between the 2006 and 2007 timetable (Data source: Peters, 2006).

resulted in travel times for 71 × 70 × 2 = 9940 relations. The results were posted in an e-mail group and are shown in Figure 9.1. It is clear that most journeys require longer travel times; also, the biggest increase in journey time is bigger than the biggest travel time reduction.

In response to these figures, NS stated that not only the travel time but also other factors play a role in customer satisfaction, such as frequency and number of required changes of trains. These could be incorporated when a 'generalised travel time' is used, and in that case there would be a net benefit for 1% of the passengers, since 11% of these generalised travel times would be significantly shorter, and 10% would be significantly longer (Veenman 12 September 2006). However, NS compared relative travel time changes, so that for instance a 4 minute shorter travel time on a 20 minutes journey (-20%) for one person would cancel out a 24 minute longer travel time on a 120 minute journey (+20%) for another person, which is also debatable. Nonetheless, the amount of passenger kilometres increased by 3 to 4% per year (Nederlandse Spoorwegen 2008; 2009a).

Another unintended effect, that easily could have been foreseen, was that the longer running times (and also longer turn-around-times at endpoints) required more rolling stock units and more staff. In 2005 and 2006, NS was cutting back on rolling stock by scrapping old types. At the same time, train staff were granted early pension. Then, during 2006, all of a sudden, huge amounts of new train staff were recruited and pensioned staff was even asked to come back. Also, old trains were no longer sent to the scrap heap, but instead were given refurbishments, and large amounts of additional rolling stock were leased from Germany and new rolling stock was ordered. Thus, NS (and via them the traveller and the tax payer) appears to have paid a high price for the increased travel times. This is ironical, since the intention of this timetable design was to lead to better utilisation in order to save costs (see section 1.3.3), and the result was the exact opposite.

In contrast is that the Swiss Railways engineered their new Bahn2000 timetable in an exactly opposite way to the Dutch approach. Where the Dutch design team in 2004 started with a fixed infrastructure and tried to fit in the train paths as well as possible, the Swiss started with a customer-oriented synchronised timetable and gradually changed the rolling stock and the infrastructure in such a way that the required running times would become feasible, which gradually resulted in more frequent services, better connections and shorter travel times (Moser 2004), which led to a growth in passenger volume of 38% in 3 years (Laube and Lüthi 2008). Most remarkable is that the Swiss were confused when Dutch colleagues asked them about their timetable strategy, because they thought that they were simply bringing the Dutch idea from the 1970's to perfection (Avelino et al. 2006).

Lessons learned

We may conclude that in the best case a three year period in which a timetable was redesigned from scratch did not lead to significant improvements with regard to the requirements mentioned in section 2.2.2. Apart from that, the higher frequencies (6 instead of 4 trains per hour) that were originally planned, were not realised. Also, when politicians realised that the strong focus on punctuality now led to longer travel times, they were quick to issue a demand for shorter travel times (Tweede Kamer der Staten-Generaal 2008a). Hence, timetable planning is not only an operational process to be left to engineers or economists (Avelino et al. 2006). In fact, the timetable (combined with the fare system) is the service that is offered to the customer and should be the starting point to determine which means of production are necessary and how these should be operated. The railways should try to produce what can be sold, rather than try to sell whatever they can easily produce. However, as was shown in section 2.2.5, there are no incentives for the railways to come up with short travel times, even

though this is one of the most important variables in determining modal split or market share. A question for the railways, the Ministry of Transport and parliament is whether this is a desired situation.

9.3 Designing and supporting the operational processes

This section will provide examples on how the means of production are designed to support the operational processes.

9.3.1 Infrastructure design

Intention
Of course, infrastructure is required to provide train paths. Although infrastructure in fact is a means of production and as such is not relevant for the public function of a railway system, it is normally financed by the government so that it is subject to public debate. Given the high price of infrastructure, a derived wish is to make sure that no infrastructure is built that does not in some way contribute to producing high-quality services. This case describes how the infrastructure design process helps to deliver train paths in relation to the functions of timetable design, traffic control, and train driving.

Practical use
When discussing this topic it becomes clear that many actors are involved in infrastructure design, leading to fuzzy requirements, long discussions and missed opportunities. For instance, due to the long time between design and actual operation, normally a timetable is not known at the design stage, so that infrastructure is designed based on a notional (planned) number of trains per hour between different pairs of stations. This means that the actual (absolute and relative) timing of these trains, which still is one of the most important elements of railway operations, is not taken into account. This makes it difficult to decide on the right places for facilities such as passing tracks, level-free junctions, or even the required number of tracks in general.

Taking advantage of large scale maintenance of the network would allow for some quick wins, since if parts of the network have to be renewed anyway, it is relatively cheap to change the design. However, because the actual train paths are not as much in the spotlight as the infrastructure elements that need maintenance are, opportunities for changes are easily missed. ProRail's infrastructure management department generally replaces old parts with identical new parts, without taking into account any possible functional improvements for the train paths that can be delivered, arguing that functional improvements are the responsibility of the capacity management department. However, the capacity management department waits for formal requests from train operating companies to enhance the infrastructure's functionality, and train operating companies focus on operating trains and not on the infrastructure on which these trains are operating – let alone that they know at what point in time there is an opportunity to enhance the infrastructure with relatively little extra costs; so that in the end, nobody is actively thinking about easily obtainable functional enhancements in train path quality.

From interviews, the conclusion can be drawn that infrastructure designers lack feedback from drivers and traffic controllers. This is acknowledged by drivers, who indicated that signals could in some situations easily be placed much more conveniently. However, drivers in general no longer report such situations, since they share the feeling that such feedback does not lead to a change. One infrastructure designer said that such feedback did exist in

former days, but when the integrated railway company was split up, this feedback was lost. He coped with that by organising his own feedback group, which was possible because he happened to know a team leader for drivers who could provide some drivers to supply feedback; but this was managed completely informally. Yet it appears to be too easy to blame just the company's unbundling for these problems; rather, this made it explicitly clear that such information exchange must be organised.

Also traffic controllers complained that the infrastructure was, according to their opinion, too much designed for the regular situation and that it has not enough possibilities to cope with disturbed traffic; e.g. by providing additional parallel sections of track where fast trains can pass slower ones. In some cases points have been removed, which decreased their ability to let trains enter and leave a station via parallel tracks and so to avoid conflicting train movements. However, ProRail's traffic control department cannot indicate which points are most desired: they ask for all possible connections between tracks, which is also of little help to prioritise which connections need to be built first, if the limited budget allows for only some of the items on the traffic controller's wish list.

Lessons learned
In terms of the cascade model, we can conclude that infrastructure design, due to its long time span for realisation and incomplete understanding of actual operation, can only provide the right conditions for real time operations at a rather rough level. The long term on which large scale infrastructure works have to be planned and the high costs involved lead to long debates, in which the effect of infrastructure modifications on timetable improvements (or train path quality) can hardly be made clear. Following the Swiss approach, such discussions could be informed by defining the desired train paths before designing the 'machine' to produce these. It is suggested to improve knowledge about how useful certain infrastructure elements are in order to deliver good and safe train paths. For current designs, this requires feedback on the actual use of these elements and a measurement of train path quality. For new designs, it requires a more detailed understanding of the intended network timetable and how operational control will cope with small deviations from it. A new set of design norms is currently under development, which appears to allow for integration of such an approach, but this has still to be operationalised (see Appendix A2).

So far, the examples have shown that the physical infrastructure itself has a relatively prominent role, both in infrastructure design and in timetable design (in which the infrastructure is roughly taken as fixed). The resulting quality of the train paths and via these, the quality of journeys over the network appears to come at the bottom of the list, rather than that it serves as a starting point for desired design modifications. It was also shown that the infrastructure design does not always optimally support the wishes of traffic controllers and drivers. Of course, it is not possible to realise all their wishes, but currently a working structure to list which wishes they have and how these could be realised, is missing.

An example, in which the train paths and the cooperation between several functions in the cascade model were meant to be taken into account, is the BB21 programme, which will be discussed next.

9.3.2 System architecture: BB21 programme

Intention
As indicated by the Nyfer study (see section 1.4), one possibility to improve performance is by applying smarter planning methods and new technology, with the goal to decrease the

probability that trains are in each other's way and create delays. Precisely this was the aim of a large innovation programme in the Dutch rail sector that started in the early 1990s. It was called BB21, an acronym that stands simultaneously for 'Beheersing en Beveiliging in de 21e eeuw' and 'Beter Benutten in de 21e eeuw' (Control and protection in the 21st century / Better utilisation in the 21st century). It was meant to enhance infrastructure utilisation by redesigning traffic management processes, supported by new, advanced technology for traffic management and train control (Poort et al. 1994). As such, it in fact meant a paradigm change: no longer should a conflict-free timetable be the only basis on which to control the process and should deviations from this plan be minimised; the new philosophy meant a constant adjustment of the plan given the actual delays, so that deviations from the original plan would now be explicitly and actively managed.

Practical use
However, this originally wide scope of the project was narrowed in 2002 by a governmental decision and ProRail indicated that this would come at the cost of the future performance (Klerk 31 March 2003). The BB21 project now was restricted to focus only on realisation of new technology, such as ETCS, GSM-R and 25 kV overhead line voltage; which were necessary for three mega projects mentioned above. Since then, the project has focussed on new technology and the process improvements have no longer been part of it.

Lessons learned
In this example, we see a clear shift from objectives to means of production. Remarkably, the final report of the BB21 project mentions a start date of only 1999 (not the early 1990s) and considers the project a success (ProRail 2008a), even though the original aims were by no means met. Yet the wish to increase network utilisation remained and now has to be accomplished without smarter (re)scheduling tools, which were considered necessary by the Nyfer study. It is to be expected that the process improvements that were the original aim of the project could still be worthwhile and should be resuscitated.

9.3.3 Working timetable design

Intention
Given its importance in railway operation, the timetable design process was already explained in section 2.3.5. In fact the working timetable is nothing different than a production schedule. As was made clear in the analyses in Chapters 7 and 8, it turned out that some possibilities for improvements can likely be found in the timetable design process. Therefore, this case will focus on whether and how the timetable design supports the operational processes of drivers and traffic controllers.

Practical use
Because the paradigm change from the BB21 project did not get developed, timetable design had to be done in the 'traditional' way when a more reliable timetable was to be designed for the year 2007. Because timetable designers rely on planning norms for running times, running time margins and buffer times between train paths, it may not come as a surprise that a new set of planning norms was designed. Timetable designers tacitly assume that trains are punctual, except for small delays for which buffer times and running time margins are required and provided. Given this assumption, they plan for green signal aspects, so that no conflicts will occur in practice. 'Otherwise, you will burden someone else with the problem', said one designer. Also, timetable designers generally believe that they do provide some space to cope with disturbances by adding spare time to the timetable. However, since this spare time can be scattered over the hour, this does not guarantee that traffic controllers can find a feasible solution to cope with disturbances: traffic controllers still complain that timetable

design is too much focussed on normal, undisturbed operations, which leaves little room to cope with disturbances.

A further complicating factor is that the amount of spare time, which is given much attention during timetable design, is not shown to traffic controllers or drivers, so that the workers who actually must decide on how to use this spare time do not know where it is, other than by experience, which has to be re-learned after every timetable change. One timetable designer indicated: 'Making a timetable is already a very difficult task. We cannot also take into consideration what drivers want'. However, providing this information could easily be automated (just as providing the timetable itself) and would not make it harder to find a conflict-free timetable. Hence, when drivers do not know where additional running time is put into the train's path, they do not know where to slow down in order to stay on the path. Thus, trains may run ahead of schedule and run into unnecessary conflicts with other trains: *'If an (…) allowance is applied immediately before a junction [in order to make the timetable more stable or make it fit at all], there is still a headway conflict remaining'* was found by simulation results (Kohls and Watson 2008), and confirmed by ProRail's performance analysis office (Weeda and Hofstra 2008) and in interviews with drivers. Despite the wish to make a conflict-free timetable, several studies have shown that the meticulously planned timetable leads to conflicts in practice. In some cases, the planned process times are not met in practice, in other cases decisions have been made during timetable design that are not optimal in practice and need to be changed by operational staff (Schotanus 2004; Weeda 2005; Vorage 2006; Weeda and Hofstra 2008).

But not only can running time margins lead to conflicts that they were meant to prevent; discrepancies between the scheduled path and actually realisable driving behaviour exist, too. Since the signalling system is not unequivocally clear, especially with regard to its yellow warning aspect, drivers must at some points drive slower than 40 km/h, but the running time calculations during timetable design do not take this extra running time into account. Hence the driver is placed in a dilemma between keeping the scheduled running times or maintaining a safe speed. This dilemma is especially mentioned by freight train drivers, who fear that their path will be cancelled by ProRail's traffic controllers if their delay grows above a certain threshold.

However, it is not only the possibility for drivers and traffic controllers to fulfil their tasks which is disregarded during timetable design. Also the behaviour of some technical designs is insufficiently represented in the timetable design process, which may render the intended timetable an infeasible production schedule and/or it may amplify delays once they occur. During interviews, the following items were mentioned which cause this to happen:

1. *Level crossings on roads near stations*: near stations delay timers are used to prevent that the crossing barriers are already closed while the train is still standing for boarding and alighting of passengers. These delay timers are calibrated to the local situation and rolling stock type. Yet when quicker rolling stock is used, the barriers are still programmed to close with a delay that fits the old rolling stock type's running times. Similar problems occur on single track lines when two trains meet: before the second train can depart towards the station where the first train came from, the barriers will open and close again, which takes time which is not catered for in the timetable design norms, and delays are built in by design.
2. *Special signals for freight trains*: freight trains may be delayed when they meet so-called L/H-signals. These are used to stop heavy freight trains before steep slopes, rather than on them, in order to prevent them becoming stranded if they come to a halt on the slope. The train may only continue when it has a free route to the top of the slope. This leads to

longer required headways, so that the planned standard headway may be infeasible in some cases. Also, an initial delay of one train may amplify itself if this is passed on to a freight train which then has to re-accelerate, which can take some minutes.

It is also remarkable that during timetable design several important checks are performed, for instance to plan in such a way that the best possible signal aspects are displayed (so that unnecessary speed reductions are prevented and also no departures on yellow will take place), or to check whether trains fit along their scheduled platforms. This is given much importance during timetable design but the same checks are hardly performed when they make most sense, which is during operations. It may introduce quality and safety risks if the checks are only met in the design, but not in practice: for instance if trains are longer than their planned platform tracks or if they have to slow down for an unscheduled stop.

Lessons learned
This example shows that even when the timetable design fulfils the norms, it does not take several small, but important details in other processes into account. The railway's control philosophy always has been such that the timetable provides the reference for operations, and deviations should be suppressed as much as possible. But should not the timetable design lead to a production schedule which helps to control the real time processes? In that case, it is required that timetable designers and their tools have a richer understanding of what will be possible in practice and what will not. Apart from a richer knowledge of how technical and human system elements behave in the dynamics of day-to-day operations, it also requires feedback from operations to update this knowledge and see whether solutions do work as intended. If the quality of service is to improve, then it is a first prerequisite to know the causes of delays. Currently, this is hardly the case, since only 5% of such knock-on delays have their cause(s) registered in ProRail's monitoring system, as a case-study for a busy line showed (Weeda et al. 2006). ProRail's new KPI 'Train delay minutes' might be helpful here, since it allows one to attribute causes to every delay minute (which punctuality does not allow for).

Especially with increasing network utilisation the current, top-down control philosophy appears to be more and more a utopian wish rather than a practical preparation for operations, in which disturbances do occur and make it necessary to take these devils-in-the-details into account. Active management of timetable deviations becomes not just a nice add-on but a critical element to ensure efficient and safe operations. The next section describes how an essential element of the traffic control system related to this has not been introduced.

9.3.4 Traffic control and the VPT control system

Intention
As was described in sections 2.3.5 to 2.3.7, the VPT system (Vervoer Per Trein = Transport By Train) is an integrated control system to support the processes of timetable design, rescheduling and route setting. The VPT project started in the 1980s and its first version was rolled out in the mid-90s. Since this system is meant to support the key operational processes, its design is clearly interwoven with the organisation in which it has to function. However, we will see that both the system's design and the organisation for which it was designed were changed, and not necessarily in a synchronised way.

Most information in this section stems from elaborate interviews with two of the system's designers who were involved from the very beginning of the project until a few years ago. They indicated that there used to be a lot of (mis)communication between traffic controllers

when there were service disruptions. Decisions were often taken in a hurry and in some cases this led to unforeseen consequences which made the troubles even worse. The information flows had to be streamlined. The goal of the VPT system was to support the traffic controller in an uncertain process, so it was meant to give a quality improvement. The idea behind VPT was that if all operators would continuously update the plan, all other actors (but not the driver – added by author) could immediately see the changes on their computer screens so that the need for communication by telephone would be significantly reduced.

In the preparations for the railway's 'unbundling' and some years before the introduction of the VPT system in 1993, Hens Biekart, the later director of NS Verkeersleiding (Traffic Control), stated that there was considerable performance variety between the seventeen traffic control centres. Although it was certainly a goal to standardise traffic control's quality, it was considered even more important to be able to answer questions with regard to why certain decisions were taken in a certain manner: with competing train operating companies that would be allowed access to the tracks, such transparency was required. In a magazine article, Biekart said that 'rules and contracts cannot replace intelligent thinking. Otherwise, we could let computers do all the work'. He also indicated that rolling stock and train crew should be dispatched by their own organisations from a principled point of view, yet he stressed that this was not yet possible from a practical point of view: *'in the case of disruptions it is necessary to be able to act as quickly as possible: cut the knots and take measures immediately (…) Then, it is useful to have the authority at one point to decide (…). If it is necessary to contact all kinds of organisations by phone, that will be disastrous'* (Nederlandse Spoorwegen 1993). It is somewhat ironic to conclude that what happened is exactly what Biekart feared most: communication times exploded since many telephone conversations are now required, and standard rules are used to regulate the traffic process, as was described in Chapter 8. This makes the system more vulnerable for breakdowns, as is described by Steenhuisen and De Bruijne (2009).

Practical use
Unbundling the railways was an unwelcome surprise, since the VPT system was intended as an integrated control system. For instance, an interface to the rolling stock circulation program (Materieelbalans) was part of the plan, yet after the unbundling, no one was really interested in that interface; also, NS had enough difficulties to keep track of the actual train compositions so that the information in the system was not always up to date anyway.

Traffic controllers now indeed work by changing the plan, yet on a less abstract level than was intended. They try to keep track of what the automatic route setting system will and will not do, which is understandable since they need to change its plan lines in the case of route conflicts. Thus, they still think in terms of routes to be set. It was intended, however, that the traffic controller would operate another module of VPT, for conflict detection and decision support (in Dutch: ConflictSignalering en BeslissingsOndersteuning, CS/BO; Lenior et al. 2006), but that module was never realised, even though successful pilots were performed (NS Verkeersleiding 2000; Stolk et al. 2000; Makkinga et al. 2002). No clear reason could be found for not further continuing this development. This module would have calculated where and when conflicting infrastructure claims would arise and it would have proposed solutions, which the traffic controller could then choose. The decision support module would then have updated the plan lines in the automatic route setting system, which would execute them whenever it was appropriate to do so.

Therefore, the automatic route setting system was not designed to have a human interface and it was also designed as a 'stupid' module that would just do what was programmed into its

plan lines. Later, several functions that were originally intended for CS/BO were added to the automatic route setting system, but the overall concept was lost and the automatic route setting system has become a 'misshapen monster' as it was called by one of the designers. For instance, the automatic route setting system was meant to only set plan lines as a whole, but, since the CS/BO module would have constantly changed plan lines according to the traffic situation, this would be fine. Now that there is no CS/BO module, this became a hurdle since several blocks could only be released at once, which in fact is a severe capacity reduction that leads to much longer headways in stations. Now, the automatic route setting system has been modified so that it also has the ability to set routes which are only part of one plan line, but the system only checks to a limited extent whether such a partial route makes sense. With the CS/BO module, it would have been known when and where conflicts would arise and routes could thus have been set until that point. Since this information is not known in the current design, the automatic system now sets routes only piece by piece, so that infrastructure elements remain available as long as possible. Thus, drivers encounter more yellow signal aspects than is strictly necessary.

Lessons learned
We see that the support systems no longer fit the current, unbundled organisation. Control processes take too much time and thus, unbundling makes the process uncontrollable. Although reversing the unbundling is not a serious option, performance gains can possibly be found by improving the decision making speed. Resurrecting the cancelled conflict detection and decision support module or similar functionality appears to be essential to guide the traffic controllers' attention and to help them to oversee the consequences of their solutions; Chapter 8 showed that this is one of the weak points of the current control strategy. Although successful trials with the module were performed, it was not introduced; the reason for this decision is unknown. It is questionable whether those who decided to cancel it oversaw the consequences of their choice. As we have seen in sections 9.2.1, 9.3.3 and also in this section, smarter ways to address and cope with deviations from the timetable seem to be essential in order to be able to deliver conflict-free train paths also in intensively used networks.

For the remaining conflicts, safety-critical backup systems are required to make sure that these will not lead to an unsafe situation. The next section will describe some developments in these systems.

9.4 Design of safety systems

The cases in this section basically refer to signalling and automatic train protection, which are functions that are required to make a train stop if it is about to run beyond the limits of the route that is reserved for it. The automatic train protection systems offers a backup function for those cases where a driver exceeds the instructions given by signalling. This section will start by explaining this backup system. Knowledge of this system is necessary to understand a number of design failures in the signalling system that will be addressed later in this section.

9.4.1 Automatic Train Protection system ATB-EG

Intention
After a serious train accident near Harmelen in 1962, it became clear that if a driver missed a distant signal this could have severe consequences, and the Ministry of Transport required NS to install an automatic train protection system to prevent such accidents from happening again. Although the technically not very complicated German Indusi system was favoured by the ministry, NS considered that this system would reduce the line capacity. NS was at that

time developing a cab signalling system, such that lineside signals could eventually be replaced by it and a capacity gain could be achieved (Jongerius 1993). Hence NS proposed to use their system-in-development also for automatic train protection. Automatic train protection was, and is, generally considered as the most effective way to increase safety levels, since it provides a technical backup for the driver. Although the conclusion of this thesis is that other measures are required in addition, the importance of well-designed safety systems cannot be denied.

Practical use

ATB-EG (Automatische TreinBeinvoloeding- Eerste Generatie; in English: Automatic Train Protection system – First Generation) is the standard Dutch train protection system. It is installed on most parts of the network. The system is comprehensively described by Apel (1968). The trackside system transmits the signalling information directly from the signal to the train via coded electrical currents through the rails. Thus, the actual supervised speed is continuously sent to the train and is displayed in the cab. A revoked route or a route extension is therefore immediately shown, also when the train is in between two signals. However, the set of displayed and supervised speeds is smaller than that of the signalling system. ATB-EG cannot indicate a full stop (0 km/h), but it knows intermediate speed steps of 40, 60, 80, 130 km/h, which correspond to the most used speed limits at various types of diverging points. Similar to a green signal aspect, the speed step 'green' implies that no other restrictions than the line's or the train's maximum speed apply (Nederlandse Spoorwegen 1989; Railned 2002). In later legislation, the speed step 'green' was wrongfully coded as a fixed speed limit of 140 km/h (*Regeling Keuring Spoorvoertuigen* 2004), probably because this was for a long time the maximum speed for most lines and rolling stock types. However, currently, more and more lines and rolling stock types are suitable for 160 km/h, but ATB currently does not allow this speed, even if it is allowed by lineside signalling. This is an inconsistency in the rules.

Since ATB-EG does not have information for 0 km/h, passing a signal at danger cannot be detected by the system. Basically, ATB-EG has the following logic to prevent trains from overspeeding (see Figure 9.2):

- If the supervision speed changes, a gong-stroke will sound to make the driver aware of the change. The change will be displayed in the cab.
- If the actual speed is lower than the supervision speed, ATB-EG is already satisfied.
- If the actual speed is higher than the supervision speed, ATB-EG is satisfied as long as the brakes are applied.
- If the brakes are not applied and the speed is too high, a bell will ring for some seconds to warn the driver and give him the possibility to apply the brakes.
- If the driver does not apply the brakes quickly enough, ATB-EG will initiate an emergency stop.
- ATB-EG has no means of knowing where a red signal is located. To cope with that, a so-called ATB cancel function was installed to remind the driver of the possibility of a red signal ahead. It was active when the following conditions were met: (1) the supervision speed equals 40 km/h; (2) the train's speed is 40 km/h or lower; (3) the brakes are not applied. In that case, a buzzer sounded to remind the driver that a red signal may follow and the second part of the braking curve (see Figure 9.2) still has to be executed. This buzzer has to be cancelled by the driver, as otherwise the train will be stopped. This ATB function was removed between 1997 and 2003; see section 9.4.2.
- A patch to the ATB-EG system, called ATB-VV (ATB Verbeterde Versie - Improved Version), was recently installed at approximately 1150 signals so that trains that are also

fitted with ATB-VV will now be stopped when they are about to pass a signal at danger. (see section 9.4.3 and Nederlandse Spoorwegen et al. 2006).

There are four main drawbacks to ATB-EG:

- Although its cab signalling was one of the main reasons to introduce the system, current knowledge indicates that the information exchange via the coded track circuits is not 100% reliable, so that it is not allowed to use the cab signal display to increase speed.
- Since ATB-EG only knows the target speed, but not the target distance, the driver should reduce speed immediately when a speed reduction sign is passed. Although this initially errs on the safe side, it leads to extra running time since trains have to run with the slower target speed of (often) 40 km/h over a long distance. Drivers may be tempted to compensate for this by braking more gently.
- ATB-EG cannot identify whether the train's deceleration rate is strong enough to achieve the target speed in time. Although sources from the system's design period in the 1960s say that the driver should apply the brakes quite hard (Middelraad 2000), there was a recent SPAD near Harmelen in which a train that complied with the ATB braking criterion decelerated with only 0,12 m/s², passing the signal at danger with 104 km/h (Inspectie Verkeer en Waterstaat 2008e).
- Unless the ATB-VV patch is installed in both the train and the infrastructure, ATB-EG cannot prevent trains from running through a red signal, as it does not know where the red signal is.

Common misconceptions about ATB-EG

The fact that ATB cannot supervise every arbitrary speed limit is also seen as a serious shortcoming by some people. Especially the large gap between the speed steps of 80 and 130 km/h is feared (by office staff, not by drivers), since it means that the supervision of a fixed 90 km/h speed restriction will only come into effect if the train's speed exceeds 130 km/h. Although various proposals have been made to fix this, no (Dutch) accidents are known that could have been prevented if supervision of this 90 km/h fixed speed restriction had been operative. The question is which problem this solution really solves.

Many people, including those inside railway companies, are surprised to hear that ATB-EG does not intervene if trains run through a signal at danger, as they obviously expect that it is the system's objective to prevent precisely that. It was a design choice (and thus an accepted

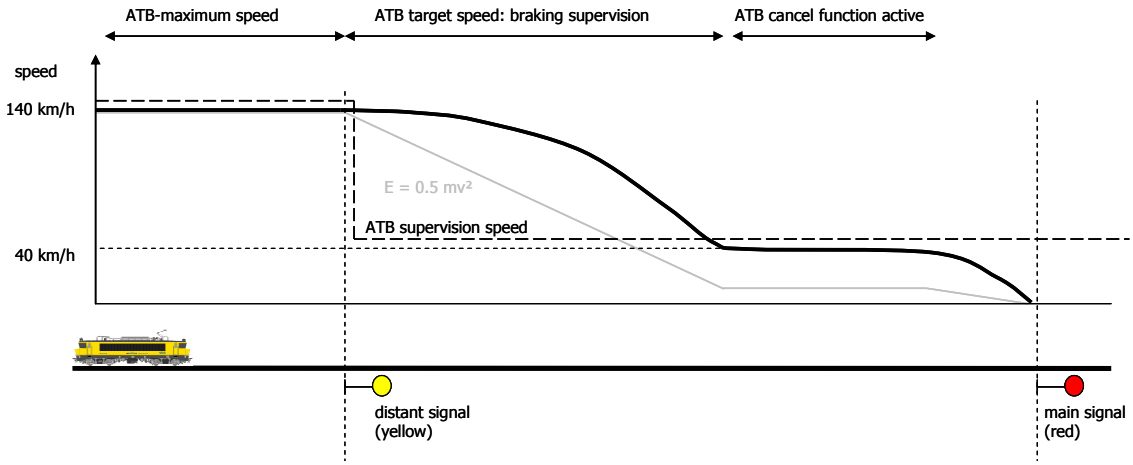


Figure 9.2: ATB-EG supervision in approach to a red signal, with an indication of the kinetic energy the train carries.

risk) that this should not be the case and, hence, not all accidents could be prevented. As was mentioned in Chapter 7, it is of utmost importance that a driver reacts to a restrictive distant signal – and this is exactly what ATB-EG does supervise, so that the speed of a SPADing train is normally limited to 40 km/h[13]. From Figure 9.2, it can be derived that the kinetic energy that is involved in a train crash at 40 km/h is much lower than at 140 km/h, so that the potential for damage is also only a fraction compared to that of a 140 km/h crash. Preventing running through red signals would have required enormous additional investments in stations, so that the additional costs were not considered to outweigh the benefits (given the technology, price levels and traffic intensity of the 1960s). The system, although it has some drawbacks, certainly improved the safety level on the Dutch rail network.

How the effectiveness of ATB-EG may be decreasing

Nevertheless, this achievement should not be taken for granted. The question should not be whether ATB-EG (or any other system) is a good system per se or not, the question is whether it helps to influence accident mechanisms in such a way that accidents become less likely. The situation on the Dutch rail network has been changing over recent decades, and it will change further. This means that also the accident mechanisms are changing over time, and therefore also the way in which ATB interacts with them is not constant. At the time of the system's design in the 1960s, most stations allowed for a maximum speed of 40 km/h for all trains. Station layouts have since been improved more and more such that through trains can pass with higher speeds of up to 160 km/h (carrying 16 times as much energy). Although the speed of a SPADing train itself is limited to 40 km/h it can enter a route in which a train is approaching with a much higher speed (Raad voor de Transportveiligheid 2001), so that the risks have gradually increased. Also train frequencies have been increased, thus making it more likely that a train is present when another one runs through a red signal.

Another reason why the effectiveness of ATB may be decreasing is that today's drivers are getting used to the system. In former times, the system was installed only on the most important lines, thus offering the driver additional protection. On the not so busy, mostly regional lines, drivers were running without ATB and still realised the possibility and disastrous consequences of missing a signal. Their driving style clearly was based on the information provided by the signals and they used ATB-EG (on those lines where it was installed) as a backup system, as it was intended. Thus, a driver in the 1960s probably interpreted the braking bell as a welcome but worrying message that told him that he had, unfortunately, missed a signal. The message was, so to speak, 'without this warning, you might have been dead in a few seconds' – a message that should be prevented in the first place and which required immediate action in the case where it did sound.

For today's drivers, this is different. ATB-EG is installed on all lines and on all trains, so drivers always have the 'luxury' of a reminder, which makes them feel more safe as they know that if they fail to notice a signal, there is a backup. This will inevitably make them less alert, based on the risk homeostasis theory (Wilde 1994). The braking bell can be heard on almost any train run in the Netherlands. Its message now will be interpreted as 'you really must brake now, even if you don't want to, otherwise I will stop the train' – which is completely different from its interpretation forty years ago. Although ATB-EG strongly reduced the probability that drivers would pass a signal without noticing it, and thus certainly has improved safety, the initial effect will now be decreasing because drivers are now getting used to the system. Today's drivers are encountering yellow signals more frequently, due to

[13] Apart from exceptions, for instance if the driver does brake, but not hard enough, such as in the 2007 Harmelen SPAD.

higher network load, and they are pushed to run as punctually as possible. This means that they want to postpone braking as long as possible; also they know that their trains have a good braking performance so that they could brake later than where the signal is located[14]. Thus, when drivers feel that ATB forces them to brake unnecessarily early, they may be tempted to brake as late as possible – that is, waiting for the bell to ring – and then brake not too strongly, in order to prevent running long distances with low speeds, which causes delays. Drivers may do so especially if they know that the signals are spaced far apart (which can be the case due to integration of main and distant signals), if they are already running at a relatively low speed or if they expect a signal clearance. Drivers indicate that if they want to make up time, this can best be done by knowing how to brake: the performance optimum is on the boundary of controllable operation.

The fact that drivers do not use ATB as a mere backup system, but have adopted it as a normal part of their work, is confirmed by the fact that today's drivers are no longer used to driving without ATB protection. In an incident in 2004 a driver had to shut down the ATB equipment due to a technical failure. He felt uncomfortable about that, was not relaxed and ran through the next red signal he approached (Inspectie Verkeer en Waterstaat 2004c). Also one of the interviewed drivers told that he had one experience where he felt really unsafe when he pulled away from a signal. The reason was that he could not remember having heard the gong stroke that normally indicates a signal clearance. Afterwards, this turned out to be normal system functioning given the station's layout, but it clearly indicates that drivers rely (at least implicitly) on the presence of the acoustic ATB cab signalling. In interviews, drivers mentioned colleagues who even rely on the braking bell to sound rather than responding to signals, behaviour that can be dangerous because ATB has only a limited number of speed steps it can handle, which means that some signal aspects are not accompanied by the sounding of the gong and the bell. If the driver relies on ATB, problem detection and part of diagnosing new information are taken over by ATB. In that case, it is required to instil pride in beating the ATB, or perhaps make its interventions unpleasant for the driver, for instance by monitoring its interventions and requiring the driver to explain these to his manager if these occur too often. ATB can also confirm a false expectation if a driver must this time brake to a full stop where he normally must only reduce his speed; this was identified as a tricky situation for drivers in section 7.3.2. In both cases ATB reacts in the same way (a gong stroke and if necessary the bell) and will thus fit the driver's expectation, thereby confirming it, regardless of whether the driver's expectation was correct or not. ATB will in this case neither correct a possibly wrong situational awareness of the driver, nor will it initiate emergency braking, because the driver has already applied the brakes himself.

Relying on ATB gives rise to problems especially in station areas, where the maximum speed is often limited to 40 km/h. In that case ATB cannot distinguish between different signal aspects: they all are indicated with the same code for 40 km/h, regardless of whether the signal aspect is yellow, green or even red. Therefore, no gong or bell sounds to reflect a passage of a yellow signal or an update of the movement authority. A driver has less ATB support in stations than on running lines; yet stations are exactly the places where the yellow signal aspect is also not as clear as it is on the running lines. It is one of the 'ironies of automation' that tasks that easily can be automated have been automated, leaving (as in this case) a random set of the more difficult tasks to the human operator (Bainbridge 1987).

[14] As an example of the improved braking performance, the 104 km/h SPAD train mentioned earlier managed to stop 'only' 226 m beyond the signal (the required braking distance from 140 km/h would be approximately 500 m), whilst the yellow signal was placed to allow 1348 m braking distance before the red signal.

Lessons learned
The examples above make clear that the intended use of ATB as a backup system is no longer the case in current operations. Although the system did make rail traffic much safer, the initial effect appears to be gradually decreasing. ATB supports drivers most of the time in their task so that today's drivers have explicitly or implicitly incorporated the system into their normal behaviour. Relying on ATB clearly can lead to dangerous situations, and is therefore prohibited. However, such a prohibition is not a fruitful approach, as it does not take away the mechanisms that lead to this behaviour: if ATB gives drivers the possibility to ease their tasks, it will inevitably be used for that. Especially because the system provides information to the driver (both visual and acoustical) it will change the driver's situation awareness, no matter whether this is prohibited or not. Prohibiting this tries to adapt driver behaviour to the technology, instead of using technology to support the driver in his task. Although it is known that drivers rely on the ATB system, there is no structural knowledge of driver behaviour (e.g. based on monitoring the juridical recording unit), be it for individuals or for the driver population as a whole. Such analyses would be necessary to investigate this phenomenon in further detail.

Another example in which driver behaviour was not completely understood, and which led likely to a decreased effect of the ATB equipment, will be described in the following section.

9.4.2 Deactivating the ATB cancel function

Intention
We have seen from section 7.3.3 that the yellow aspect in fact gives a double instruction: slow down to 40 km/h *and* be prepared to stop. By its technical design and earlier design choices the in-train ATB equipment could only supervise whether the driver complies with the first instruction, that is, whether he reduces speed to a maximum of 40 km/h. For the remaining part of the route, it is the driver's sole responsibility to check the aspect of the next signal. The train protection system ATB was therefore originally developed with a so-called cancel function (in Dutch: kwiteerfunctie) that was meant to fill the gap in supervision that came into existence once the train reached a speed of 40 km/h. If the train's brakes were not applied, a buzzer would sound every 20 seconds and the driver had to push a button to acknowledge that he was still aware of approaching a possibly red signal; if he failed to do so, the train would be stopped automatically. Thus, the function was meant to focus the driver's attention onto an upcoming stop signal.

Practical use
A first error is found in the regulations (Nederlandse Spoorwegen 1989). They indicated that if the buzzer sounds, the driver must push the cancel button. This however is not an end in itself, so that it would have been better to prescribe that if the buzzer sounds, the driver should check the signal aspect and apply the brakes if necessary; if this is not necessary, he should push the cancel button. The function's meaning was clearly to make sure that the driver was still in control of the situation, and not to check his ability to push a button. However, there were some alleged drawbacks:

1. The function was also active when the train had a through route with a reduced speed of 40 km/h, so the driver had to push the button also in some cases when no red signal was being approached. Thus, there is not a 1:1 relationship between the buzzer and a red signal aspect.
2. Pushing the button in response to the buzzer became a routine action, so that the alerting effect was said to be lost.

Thus, it was assumed that the function was not effective and it was decided to remove it from all trains (van den Hout 1992), which took place between 1997 and 2003. For most rolling stock types, the removal was finished in 2000. Note that this is exactly the same period in which the rise in SPAD numbers took place – which alone is not a proof of a causal relationship, but it begs for another check of whether the function's removal was correctly underpinned, so as to be able to exclude that this is a possible cause of the increased SPADs.

A second logical error was made in parallel during the same period, which compounded the confusion. This occurred when a new type of dead man's button was being installed. The former type had to be pressed continuously – which was impracticable for drivers and which could easily be done by putting a heavy bag on it. It was therefore changed for an intermittent type of button, which has to be released at least once every 60 seconds[15]. Now, railway staff linked the two actually unrelated issues and concluded that there were two buttons to check the driver's alertness and one of the buttons could be removed (Holleman 1991; van den Hout 1992). So they removed the cancel function on the ATB and left the dead man's button. Sadly enough, this is incorrect logic. The function of the dead men's button is to acknowledge that the driver is still present. The function of the ATB cancel function was totally different, namely to remind the driver of the possibility of a red signal ahead, which was not covered by an automatic ATB brake intervention nor by the dead man's button. It provided an alarm, although not in a very specific manner. This function cannot be taken over by pushing the dead men's button. What could have been decided, was that the dead men's handle need not to be activated as long as the driver indicates his presence by using any other device, such as the ATB cancel button. This indeed was decided when the ATB was introduced, as archived minutes from the ATB study committee revealed (exact date unknown, approx. 1963).

Whether or not the cancel function was effective is difficult to measure, as it would require counting how many trains did *not* run through a red signal, where they would have done so if the cancel function was not installed. Of course, no data provides this kind of information. However, reversing the argument, as has been done by the railways, also does not work. They stated that the function was not effective because trains still ran through red lights, even though their drivers were pushing the cancel button - an argument that is analogue to reasoning that seat belts in cars are not effective because there are still casualties in road traffic, so that seat belts can be abolished. By this design change, the cancel function that became active when a driver approached a red signal (and in a number of other scenarios, giving false alarms as described) and was believed to be ineffective because pushing the button became a routine task, has now been replaced by a function that the driver has to use all day long. This function does not distinguish at all between the normal situation and that in which a driver is approaching a signal at stop.

The system as it was, was certainly not perfect. But the system may still have served as a reminder in those cases where a driver was really approaching a red signal, and it should not have been taken away. By removing the cancel function, its effectiveness will certainly not improve. At least one of the interviewed drivers indicated that he had used the function as a reminder. If the buzzer sounded, it indicated that he was not really aware and he therefore always tried to cancel the function before the buzzer came. Certain improvements, aimed at giving warnings that are more strongly related to the driver's task, should have been

[15] Sometimes it is wrongfully believed that this type of button only intervenes after 60 seconds if a driver becomes incapable, but this is only the case if the driver keeps the button pushed. If he releases it, which is likely when his muscles relax, the system will intervene after a few seconds.

Table 9.1: SPAD data before and after removal of the ATB cancel function.

	Relative number of SPADs			Absolute number of SPADs		
Rolling Stock Type	1994-1996 (with)	2001-2003 (without)	apparent increase	1994-1996 (with)	2001-2003 (without)	actual increase
SGM	11,2%	6,7%	-40,2%	51	51	0%
ICM	15,0%	11,0%	-26,7%	69	84	23%
E1700	2,7%	1,8%	-33,3%	12	14	12%
Mat'64	17,2%	18,8%[1]	9,3%	79	144	83%
Sub total	46,1%	38,3%	n/a	211	294	39%
Old rolling stock	53.9%	-	n/a	247	-	n/a
New rolling stock	-	61,7%	n/a	-	473	n/a
Total # SPADs	100%	100%	0%	458	767	67%

[1)] 2004 data only

considered. One could think of warning only at braking distance from the red signal, although this could also make drivers complacent.

Because the number of SPADs was increasing at the same time that the cancel function was deactivated, also the Dutch parliament asked questions about this case (Tweede Kamer der Staten-Generaal 2005b). In the way that these questions were answered, again a reasoning error was made. A table was provided in which SPAD data before and after removal of the ATB cancel function were compared. This table corresponds to the gray area in Table 9.1. I constructed the remainder of this table by multiplying the percentages with actual SPAD numbers of the corresponding periods. These were 458 in 1994-1996 and 767 in 2001-2003 (Griffioen and Vorderegger 2001; Inspectie Verkeer en Waterstaat 2006).

From the percentages in the grey area it appears as if less SPADs occurred once the function was removed, and the railway inspectorate, who answered the questions, concluded that removal could not be a cause of increasing SPADs. However, the figures shown in the table are confusing, if not even misleading, so that the conclusion of the inspectorate is neither justified nor refuted. In order to make a sound statement about the effectiveness of the cancel function, one should look at how well it succeeds in decreasing the SPAD rate per approach; so the number of red signal approaches must be known, and also how often this leads to a SPAD. Also, all other factors must be kept the same (ceteris paribus requirement). The table does not satisfy these criteria at all.

1. *The percentages do not indicate SPAD probabilities per approach*, but how the SPADs are distributed over rolling stock types. The table can only include rolling stock types that were in use both before and after the function's removal. Since the rolling stock fleet was increased over the period, it is not surprising that the older rolling stock types now have a lower share in SPAD incidents (46,1% to 38,3%). However, this says nothing about the effect of the removed cancel function, especially because the total number of SPADs increased by 67% as is shown in the right-hand-side of Table 9.1. Although the percentages in the grey part of the table seem to indicate a decline, they in fact represent a rise in SPAD numbers. Instead of percentages of the total number of SPADs, percentages of the total number of red signal approaches should have been used. Since these are not monitored, I (and the inspectorate) could not produce such data.
2. *The ceteris paribus requirement is not met*; because the comparison is made in two different time intervals, it is not clear whether any changes must be attributed to the removal of the cancel function or to other changes which took place in the same period. Instead, an experiment with a control group should have been carried out.

3. *The table is confusing*. The numbers are not weighted according to the rolling stock use. Incorrectly, this gives the impression that a 1% increase of SPADs at one rolling stock type balances a 1% decrease in another type. This is only true if both rolling stock types encountered red signals equally often, which need not be the case, for example because of differences in fleet size or in service types (intercity/local train).
4. *The selection of rolling stock types is incorrect and not transparent.* Of course, only rolling stock types can be compared that were in use with and without the cancel function. In that case, the 1600 and 6400 series could have been included as well, and the 1700 series was the first to come into service without the cancel function and is included in the table mistakenly.
5. *The effect of new rolling stock is not clear from the table.* The total increase in SPADs was 67%, but the (old) rolling stock types included in the table only account for a 39% increase. Not only does this 39% increase already invalidate the conclusion that there is no increase in SPADs due to removal of the cancel function, it also means that the new rolling stock types introduced between 1996 and 2001 must have much higher SPAD rates than the old ones: indeed, the rolling stock not included in the table accounts for 473 of 767 SPADs. These new types came into service without the cancel function. In neither case does it prove that there is a causal relation behind these changes, but it should at least ring a bell that the SPAD numbers increased much more for newer rolling stock types and went up overall.

The parliament also noticed that the data provided in the table were not sound, and asked again how the data in the table underpin the conclusion that removing the cancel function did not lead to increased numbers of SPADs (Tweede Kamer der Staten-Generaal 2005a). However, when these questions were answered, a rhetorical trick was used in which, when considered in detail, the argument of the answer traces back to the same table that caused the questions. To compound this confusion, Dutch legislation for rolling stock homologation requires trains to have the ATB cancel function installed and even explicitly forbids integration with the dead man's device (*Regeling Keuring Spoorvoertuigen* 2004). Again questions in parliament were asked, since it is remarkable that a compulsory element of the train protection system was removed, and a remarkable answer again was the result: '*The technical cancel facility is part of the standard specifications of the ATB equipment (…). [Since 1995], the technical function is still installed, but it is not activated'* (Tweede Kamer der Staten-Generaal 2008b).

A more elaborate and Dutch version of the text in this section of the thesis was published on a Dutch conference for safety scientists (Van den Top 2009). Since the railway inspectorate answered the questions of parliament and was also the source of the data provided in the grey part of Table 9.1, the concept version of this conference paper was sent to them for comments. Remarkably, the main point of their remark and one of the first things they said was that 'if this will be published, the worst thing that could happen is that parliamentary questions will be asked again'. The fact that the underlying logic is wrong was heavily denied, and the inspectorate was convinced that, even if the logic indeed were incorrect, the decision to remove the function was a good one. Similarly, when I wrote a critical article with a colleague on the incomplete findings of one of the inspectorate's accident reports (Van den Top and Groeneweg 2007), the inspectorate was not so much worried about the identified safety gaps, but quick to ask the authors to explain why they wrote such an article.

Lessons learned

The example described in this section leads to several questions that need to be addressed and lessons to be learned. In the first place, this example shows a change being made in a train

protection system on grounds which are ill founded. Although the design of the ATB cancel function had its limitations so that its effectiveness was not optimal, no clear evidence existed that the ATB cancel function was totally ineffective so that it could be removed. A number of errors in the line of reasoning that led to this decision were identified. This demonstrates that it is insufficient to think that SPADs have their cause in driver error alone: also system designers make safety-critical errors. Because the cancel function was (albeit also ill-founded) used as one of the arguments to increase the maximum speed following a yellow aspect from 30 to 40 km/h (see section 9.4.4), a complete and consistent analysis should have led to the outcome that this change also should be reversed, or at least reconsidered. We see that insufficient data were available to make a proper analysis of the system's effectiveness. Especially the need to include the exposure, the number of red signal approaches, is easily forgotten, but this is essential to perform analyses regarding the effectiveness of measures to prevent SPADs.

The analysis made by the railways was correct insofar that a function is necessary to focus the driver's attention to what is important at the moment. Drivers now have less support in their task than was originally intended: they are no longer reminded of the possibility of an upcoming red signal. The question how this function could be fulfilled in a better way was not addressed; there was a focus on the means, not on the goals that should be achieved by these means. The driver's situation awareness could also have been improved by other modifications to the system, such as making the driver only need to acknowledge at a real braking distance (ca. 300 m at 40 km/h or 200 m at 30 km/h) before the stop signal is reached, or by only sounding the buzzer but removing the button. Now, the old shoes have been thrown away before new ones were bought.

Another issue is to be raised on a political and societal level. It is a formal task of the railway inspectorate both to enforce the law, and to have an open eye for potential safety flaws. Since the Minister is responsible for rail safety, answering parliamentary questions normally means to defend the position of the Minister, rather than being open to criticism expressed in such questions. Yet it is the inspectorate which prepares the Minister's answers. Hence the inspectorate gets a double task to fulfil. It is questionable whether it is possible to fulfil both tasks at the same time.

9.4.3 Deploying ATB-VV

Intention
As we have seen in the previous section, the ATB cancel function was maybe not as effective as was desired. To combat SPADs, the sector-wide SPAD workgroup planned to deploy a patch to the ATB system, called ATB-VV (VV stands for 'verbeterde versie'; improved version), so that the system will also stop trains that approach signals at danger, as was described in section 2.2.7. It was intended to use a risk-based approach to decide on the thousand most risky signals where the patch would be installed. To define which signals met this criterion, a method was derived by ProRail and the railway inspectorate (Wright and Vorderegger 2005). To implement this method, the same database was used as in section 7.3.6 (see also Appendix D).

Practical use
Quickly after the workgroup was installed, the Ministry of Transport reported: *'to tackle the SPAD problem, two tracks are essential. First, there is the human factor: which possibilities exist from the side of operating trains to reduce the number of SPADs with the perspectives of culture, attitude and behaviour. The second track is the technical one, where a difference*

must be made between optimising the positioning and visibility of signals and on the other hand changes to the automatic train protection system' (Ministerie van Verkeer en Waterstaat 2004b). Although this appears to be a logical division, it splits the problem at its core, namely the interaction between human drivers and the technical design of the signalling system, the automatic train protection and its underlying choices. The previous sections on ATB have demonstrated the importance of taking this interaction into account, as will also the coming sections about changes to the signalling system.

ATB-VV was to be installed at the thousand most risky signals. But although ATB-VV is said to be risk-based, the first two deployment rules already contradict this principle, because it was decided to install ATB-VV:

- at all multi-SPAD signals – signals with significantly high SPAD numbers. This indicates a high SPAD probability only, but says nothing about the consequences;
- at all locations where a SPADed train could penetrate into a route where a train could pass with 100 km/h or more (but this says nothing about the probability of SPADing).

Purely mathematically speaking, both criteria do not lead to a risk-based approach, since either the probability or the consequences are unknown. In both cases, the exposure rate (number of trains approaching the signals at red) is not taken into account, which is (together with probability and consequences) one of the three factors that make up risk (see section 3.2). In order to make a statement about the effectiveness of the chosen locations, I have calculated the SPAD probabilities for these signals in the same manner as was described in Appendix D. The results are displayed in Table 9.2. It can easily be seen that the signals leading to a high-speed track have a low probability of being passed, whilst the multi-SPAD signals have, not surprisingly, a probability of being passed that is on average almost ten times higher than for an average signal, which clearly indicates that there must be difficulties for the driver to react adequately to the signal.

Table 9.2: SPAD probabilities ($\times 10^{-6}$) per approach at either one signal aspect for SPAD workgroup signals.

	average
average signal	2.9
average signal on post	1.8
multi-SPAD signal (3 or more SPADs in 5 years)	25.0
signal leading to fast track (V ≥ 100 km/h)	1.6
ATB-VV 1000 signal list	5.0

Although it is defendable to install countermeasures at multi-SPAD signals where, apparently, something is wrong, especially the second criterion leads to excessive examples of ATB-VV equipment installed at places where the SPAD likelihood is extremely close to zero, for example on tracks that are only accessible after removing a derail or on tracks that have not seen a train for the past ten years (Figure 9.3). Again, this indicates that decisions are taken on the basis of oversimplified beliefs of how the system works, instead of gaining more detailed understanding of the (sometimes failing) processes that users actually are performing. As was shown in Chapter 7, the workgroup's database did not allow for analyses of signal approach situations, yet these have the highest relation with the driver's process and thus the SPAD probability. Instead, it appears that those who decided on the locations for ATB-VV equipment tried to gain an understanding of what was going on by looking at properties of the physical elements instead of their function in the system, as if they were to assess the quality of a newspaper by looking at its paper quality instead of the quality of the article the newspaper publishes.



Figure 9.3: ATB-VV equipment installed at ineffective locations.

This is in line with the top-down control of the cascade model that was described in section 2.3.1: designers apparently try to improve on the basis of their belief about how the system works, without checking their assumptions or the results of their analysis. Also in this case, there is no feedback from operations to systems design. The SPAD workgroup expected ATB-VV to be highly effective to reduce SPAD numbers. However, when the Dutch railway inspectorate analysed how many SPADs would have occurred if the system had already been installed, it was revealed that its effectiveness is probably much lower than was expected (Inspectie Verkeer en Waterstaat 2008h). Only when the SPAD workgroup became aware of the fact that the intended measures would not lead to the desired result, was it decided to analyse multi-SPAD signals in a multi-disciplinary way with expert teams; in these analyses, the added value paid off since causes in the relationship between human, technology and process control have indeed been found as a result, as was mentioned for example in section 7.3.2.

Lessons learned
Many improvements were expected to be gained from the ATB-VV patch, yet the nature of SPAD incidents was not completely understood. An integrated approach, taking all functions of the cascade model into account, is required to understand SPADs from a process-related viewpoint.

The signal selection method was derived by ProRail and the inspectorate; now that this method appears to be leading to disappointing results, it is questionable who will be able to perform independent inspections, for instance to answer the question whether the method itself is incorrect or whether it has been incorrectly applied. Although the involvement of the inspectorate can be considered as a helpful gesture and it is surely done with positive intentions, the additional activities appear to have the potential to compromise the inspectorate's main task, which is to act as an independent supervisor and to improve knowledge about safety hazards and how to prevent them.

As we have seen in this example and also in the previous, design modifications in the ATB take the driver's task insufficiently into account. The next section will show yet another example, dating back to the 1960s, in which the interaction between signalling, ATB and the driver's processes was also not clearly taken into account.

9.4.4 The changed meaning of the yellow aspect

The changed meaning of the yellow signal aspect, the subject of this example, shows that designers and issuers of rules in some cases are not sufficiently aware of the implications of their changed rules. We will see that a reasoning error was made in the 1960s and that this is likely to lead to increasing problems nowadays. Since this example cannot be fully

understood without some historical background, the example is quite elaborate and therefore has another structure than 'intention – practical use – lessons learned' in the previous ones.

Signalling system 1946: speed regimes, route knowledge and braking distances
Around 1950, the Dutch railways had several speed regimes, that were applied not only in the signalling system 1946 but also in the older, mechanical signalling system. These regimes are now no longer explicitly applied, but they need to be understood to be able to find a reasoning error in a design change relating to the current signalling system. A driving regime was a set of coherent rules for the driver, for example:

> *Medium[16] speed:*
> *More than 45 km/h, but no more than 90 km/h, yet such that:*
> *1st the trains' braking distance on a level track is no longer than 500 m;*
> *2nd the local maximum speed will not be exceeded.*

The other regimes are comparable and summarised in Table 9.3, in which of course the local maximum speed was never allowed to be exceeded and is therefore left out of the table.

Table 9.3: Driving regimes in signalling system 1946.

Regime:	Speed range:	Braking distance:	Signal aspect NS'46:
high speed	> 90 km/h;	< 1000 m	high green
medium speed:	45 – 90 km/h;	< 500 m	centre green
low speed	< 45 km/h;	< 250 m	low green
restricted speed	< 30 km/h;	< sight distance	yellow / flashing yellow 180
(short braking distance)	(not applicable)	(continue to brake)	flashing yellow 75
Stop	0		red

The regimes high, medium and low speed were used in cases where variable speed restrictions were in place, for example at junctions or if a train would run via a side track. They were indicated with a green aspect, where the position of the aspect in the three-lamp signal indicated the speed. When the topmost lamp showed green, this meant high speed; the other two lamps showed white so that the driver could establish where the green lamp was relative to the other lamps. A disadvantage was that route knowledge was required to interpret the signal aspects correctly: if a driver approached a junction that could be run over with 70 km/h, the signal aspect would show medium speed, and the driver should know that this meant 70 km/h at this particular spot.

The yellow aspect was used to announce a stop and provided a full braking distance until the stop signal.

> *Decrease or regulate speed such that the train can be brought to a halt before the next signal that can show 'stop' (distance at least 1000 m).*

In some cases, another signal was present within this braking distance. Although it had no function in this situation, it of course had one in other cases (e.g. short block sequences to enhance capacity). Such a signal would then show a flashing yellow aspect (75 flashes/min), meaning 'short braking distance'. For the driver, this indicated that no different regime applied after passing the signal than the one before, so that he should continue to slow down. The regime 'restricted speed' was used in cases where a train was sent to an occupied track and was indicated with another flashing yellow aspect (180 fl./min).

[16] In the original wording from the US supplier of the system, this was called mediocre speed.

In conclusion, all green aspects indicated that a driver could proceed, albeit in some cases with reduced speed, whilst all yellow aspects meant that the driver must prepare to stop. The signalling system 1946 was clear with regard to the available braking distances, but not with regard to the permissible maximum speeds.

Signalling system 1954: speed signalling system
The 1946 system was soon found to be error-prone, and NS wanted to replace it by another system, the signalling system 1954. The following reasons were found (N.V. Nederlandsche Spoorwegen 1968):

- During operations, the indication of high, medium and low speeds was error-prone since the position of the green lamp(s) was not always clear and confusion with white lamps was not unlikely;
- The designers wanted to get rid of the need for route knowledge to interpret the aspects for high, medium and low speed;
- The designers wanted to achieve shorter headways, which would be possible if trains could run at the highest possible speeds. This required a more precise indicated of the maximum speed.

In 1956, the new signalling system 1954 was introduced, and this is still in use today. It can display maximum speeds in steps of 10 km/h by indicating a number with the aspect. To keep the terms from the previous signalling system consistent with the new one, the definition for 'low speed' was changed and was reduced to 30 km/h – the same level as was also in use for 'restricted speed'. Although one of their constraints was now the same, both terms were still to be used in different situations: low speed for driving on at a reduced speed (e.g. aspects green and flashing green), and restricted speed if there is a need to stop (e.g. aspects yellow and flashing yellow).

As was shown in Chapter 7, the current signalling system is designed such that braking distances are implicitly met by reducing the train's speed such that the speed always 'fits' the distance between two successive signals. At short signal distances, this results in the yellow-yellow-red aspect sequence. The available aspect 'short braking distance' which was available in the previous signalling system, was considered not to be in line with a speed signalling system and had therefore been abandoned and replaced by a normal yellow aspect (Middelraad 2000). Thus, the yellow aspect – in contrast to the former signalling system – no longer gives information about the braking distance behind the signal, nor does it indicate the maximum speed it can be passed with. Because of the limited amount of information the yellow aspect provides, it is important that the driver can react to the red signal itself. Indeed, signals should be placed in such a way, as is shown by company-internal regulations from the former, unbundled Dutch railways, which quote the railway law 1956, reprint 1973 (N.V. Nederlandsche Spoorwegen 1973):

> *Article 17, paragraph 2:*
> *The aspect of signals applying for trains must under normal weather conditions be visible from the train at such a distance, that the train can be stopped before the signal that shows 'stop'.*

This means that the visibility distance must be as least as long as the braking distance from the (reduced) approach speed, including time for the driver to react. The intention of the current signalling system was that a driver, who encountered a yellow aspect, immediately should reduce speed to 30 km/h and continue by driving on sight, such that the train could be stopped within sight distance of a stop signal:

Reduce speed to 30 km/h. Continue to brake until the train can be brought to a halt within sighting distance of a signal that displays 'stop'.

Literal translations of the changes made to the meaning of the yellow aspect are given in Appendix C. Although the speed regimes had lost importance in the new system, especially because the difference between low, medium and high speed could now be made explicit, some of them were still used. In 1963, their definition from the signalling rule book read (Studiecommissie ATB 5 juli 1963):

Low speed: unless locally another speed is prescribed, 30 km/h, yet with the restriction that the braking distance on level track must be shorter than 250 m.
Restricted speed: at maximum 30 km/h, with the restriction that the speed should be regulated such that, with regard to the visibility conditions, the train can be stopped in due time with respect to the danger that must be expected.

Notice the details: low speed allows the driver to run 30 km/h and provides at least 250 m of braking distance; restricted speed does not do so and therefore allows driving 30 km/h *at maximum.* The yellow aspect thus was defined in terms of the *restricted* speed: 'the driver should, when passing the yellow signal, prepare to run with restricted speed, where the red signal is the first point of danger' (Studiecommissie ATB 23 januari 1964). This is an important notion for the remainder of this example.

Table 9.4 summarises how the new signal aspects relate to the speed regimes, and how this has changed through the years. Not only the notion 'short braking distance' but also 'restricted speed' has now gradually been abandoned. In conclusion, the signalling system 1954 is clear with regard to permitted speeds, but not with regard to available braking distances. Therefore, a driver who passed a yellow aspect was initially instructed to drive on sight as soon as he had decreased speed to 30 km/h. Also, it was prescribed that signals must be visible from the required braking distance.

Table 9.4: Meanings of signal aspects in the signalling system NS'54.

Regime:	Aspect NS'46	Aspect NS'54 (1956)	Aspect NS'54 (2009)
'no restrictions'	high green	green	green
'reduced speed'	centre green	green + number	flashing green + number
low speed	low green	flashing green / low green	flashing green / low green / (unintentionally: yellow / flashing yellow)
restricted speed	yellow / flashing yellow 180	yellow / flashing yellow	(unintentionally cancelled)
(short braking distance)	flashing yellow 75	(cancelled)	(cancelled)
Stop	red	red	red

Change to the meaning of the yellow aspect in 1965
In 1965 the maximum speed for 'low speed' was raised to 40 km/h. However, this was not only done for the green aspects related to 'low speed', but also for the yellow aspect, which in fact was defined on the basis of 'restricted speed'. Although both regimes represented

something with 30 km/h, the 30 km/h for low speed was the result of maximum speeds on points, while the 30 km/h at restricted speed resulted from visibility and braking distances.

The wish to increase the speed from 30 to 40 km/h was introduced by the study committee ATB, a panel that was installed to study how the automatic train protection could best be installed in the Netherlands. They gave the following reasons to increase the speed:

- There was a wish to use ATB also to supervise speeds in sharp curves. Due to the restricted number of possible supervision speeds, the speed in some curves should be reduced (e.g. from an unsupervised 100 km/h to a supervised 80 km/h) and additional running time would thus be introduced. The committee indicated that this lost time could be earned back and '*an increase of the low speed to 40 km/h would in the calculated cases offer more than compensation for the time losses caused by the ATB straitjacket*' (Studiecommissie ATB 26 juli 1963); this was later followed by the statement that '*the increase of the low speed from 30 to 40 km/h is one of the most important running time compensating measures that could be taken when ATB is to be introduced*' (Studiecommissie ATB 22 november 1963).
- Also, availability of new points constructions allowed for speeds of up to 40 km/h in diverting direction. If this speed gain was only permitted on green signal aspects, the running time gain would be three times smaller than when it would also be permitted after passing a yellow aspect (Studiecommissie ATB 23 januari 1964).

The following 'underpinning' (italicised) was given (Studiecommissie ATB 23 januari 1964):

- *The dwarf signal implies (…) through its form the notion of 'low speed'. From a principled viewpoint, such a form of signal must have an unequivocal meaning, so either 30 or 40 km/h.*
  However, that is not true, since the meaning of a yellow and yellow flashing aspect is 'restricted speed', also if this is shown in a dwarf signal. The form of the dwarf signal only implies that no higher speed than 'low speed' can be signalled.
- *The lowest supervision speed of ATB-information is transmitted as 'no code'. If only the meaning of the green and flashing green aspect would be changed, and the meaning of yellow and flashing yellow would remain on 30 km/h, this would require an additional code for 40 km/h. The costs involved would be very high (…). Increase of the low speed from 30 to 40 km/h can from a viewpoint of costs only be realised, if it is acceptable to increase the supervision speeds associated with yellow and flashing yellow, too.*
  The inability to make a cost-effective difference between 30 and 40 km/h in the ATB system is a correct notion. What is incorrect, however, is that the formally permitted speed is increased purely because it cannot be supervised. The option to keep the restricted speed at 30 km/h and supervise it at 40 km/h was taken into account but considered as an inferior option: *'In case this option [an integral raise to 40 km/h] would lead to difficulties in discussions with the inspectorate, another option, yet a less elegant one, would be to legally keep the restricted speed at 30 km/h and supervise it at 40 km/h in the ATB equipment'* (Studiecommissie ATB 22 november 1963).
- *The maximum speed of 30 km/h is just too low for the locomotives series 1100 and can lead to burnt resistors.*
  Although this can be true, the point has no relationship with the restrictions following from diverting points and braking distances, which are the relevant restrictions for the definition of the low and the restricted speed.
- *Even though it appears as if the increase of the maximum level for the restricted speed leads to an increased safety risk, one could counter this since the maximum speed of 30 km/h is now only prescribed in regulations and is regularly exceeded in practice. The new limit of 40 km/h will be supervised by the ATB system, and moreover, the ATB cancel*

function will make sure that the driver has to acknowledge his continuing alertness. Now it may be assumed, that drivers who ran faster than allowed only did so on those spots where they knew that this could be safely done; it cannot be an argument to make such exceptions the rule. Also, the supervision of the 40 km/h limit was later shifted to 43 km/h so as to preserve drivers from many interventions. And the ATB cancel function has now been abandoned (also with a logically impaired underpinning, see section 9.4.2), so that that part of the underpinning no longer exists.

Although it was apparently acknowledged that a higher speed requires either longer visibility distance or shorter reaction times and/or shorter braking distances, the choice was made to increase the low speed , but also the restricted speed, to 40 km/h. Via several smaller changes, the current meaning of the yellow aspect (*Regeling Spoorverkeer* 2004) is now:

> *Limit speed to 40 km/h and continue, depending on visibility and braking conditions, with such a speed that it is possible to stop in front of the next signal that orders to stop.*

Incrementally, also the notion that the driver should drive on sight has vanished. Originally, the plans of the ATB study committee were to also increase the restricted speed to 40 km/h for driving on sight (i.e. flashing yellow aspect) but this was not realised until the 1990s.

In conclusion, the ATB study committee sometimes mixes up the notions of low and restricted speed. In an apparent wish to shorten running times, debatable arguments are provided to underpin an increase from 30 to 40 km/h, whilst necessary arguments in relation to the required sighting and braking distances remain unmentioned.

Consequences of the change

Whilst it must have been the case that drivers approaching a stop signal had the ability to first detect this signal, before they were required to brake, this is currently not always the case. An increase in the maximum speed from 30 km/h to 40 km/h does not only lead to longer braking distances, it also leads to a longer distance travelled during the driver's reaction time. If a driver needs 9 seconds at 30 km/h, he travels 75 m, at 40 km/h this is 100 m. If a signal is then visible on the minimum required distance of 200 m, the remaining braking distance should than be at maximum 125 m from 30 km/h or 100 m from 40 km/h. The ATB study committee (Studiecommissie ATB 22 november 1963) provided information on braking distance, from which it could be derived that freight trains at 40 km/h will not have sufficient braking distance, while they can stop from 30 km/h only by applying the emergency brakes. For passenger trains, the braking distance at 40 km/h will be sufficient, but again, only in case of emergency brake application. However, it is a design assumption of the signalling system that a driver should be able to react to signals using a normal service brake application, as is formulated in AV130 which was issued by NS in 1963 and can now be found back in ProRail's design rules (ProRail 2005a). In summary, after the change, a driver can only stop in time if one of the following situations applies:

- the signal is visible from a longer distance than required;
- the train's braking performance is better than required;
- the driver reacts faster than required;
- the train runs slower than required.

However, each of these situations will only exist by coincidence and cannot be considered a sound basis for a safety system. Also, some of these situations are increasingly under pressure, e.g. due to the design of more and more complex station layouts and due to increased pressure to operate trains punctually.

Current situation
Although the use of signals has not significantly changed since 1954 – some station layouts still stem from that time – the rules of how to interpret the signals have gradually changed. This has now led to an inconsistency between the following formal requirements:
- The railway regulations, that state that signals must be visible at braking distance:

> *Article 27 (Regeling Spoorverkeer 2004):*
> *Signals shall be placed such that the driver can, depending on the locally permitted maximum speed, perceive these in due time and react to them in an appropriate order.*

- The formal meaning of the yellow aspect, which is contained in an appendix to the railway regulations, and allows a speed of 40 km/h unless visibility or braking performance is reduced:

> *Railway regulations, Appendix 4, signal aspect 212 a/b:*
> *Limit speed to 40 km/h and continue, depending on visibility and braking conditions, with such a speed that it is possible to stop in front of the next signal that orders to stop.*

- The signal sighting requirements, which are issued by ProRail as a further specification of the legal requirements:

> *ProRail, AV133.1, article 2.4.2.1:*
> *[Sighting] distances are based on firstly a desired visibility during 9 seconds and secondly a minimum distance of 200 m such that the driver can perform the last part of his braking distance in an appropriate way.*

Note that the current indication 'depending on visibility and braking conditions' does not fully cover the original intention of 'restricted speed' or 'driving on sight'. Even in the case of perfectly clear visibility conditions and braking performance, a driver can be unexpectedly diverted to another track as an accident in Amsterdam on 21 May 2004 made clear, so that even this situation asks for a lower speed (Inspectie Verkeer en Waterstaat 2004b). If it is desired to have a more explicit formulation of the driver's behaviour in the law, this could be done as follows: *'drive with a speed such that the driver can stop within the distance over which he can oversee his route'* (Wedzinga 2002).

In conclusion, the design philosophy of the current signalling system, the former law and the current law are based on the principle that a driver should be able to react to a signal once this has become visible. Given the minimum sighting distance of 200 m, which is not easy to change, the maximum speed of 40 km/h is too high and should be reduced to 30 km/h for passenger trains and approx. 20 km/h for freight trains. The current maximum speed of 40 km/h is based on a reasoning error and makes rail safety of secondary importance to running time gains. At this speed, it cannot be guaranteed that a driver can stop in front of a signal that becomes visible at the shortest possible visibility distance. If it is important to achieve this running time gain, other measures, such as improved braking performance or longer minimal sighting distances, or better distant signalling should be taken so as to make the higher speeds possible in a safer way.

Additional implications: the system has become more and more vulnerable to this error
There are several reasons that indicate that the railway system has become more vulnerable to this error. These will be elucidated below and are:

- Signal aspect 'flashing yellow' (drive on sight);
- Education and route knowledge;
- Increasingly complex station layouts with shorter sighting distances;
- Automatic route setting.

Signal aspect 'flashing yellow'
The speed increase for the yellow aspect took place in the 1960s. The original intention was to increase the maximum speed also after a flashing yellow aspect ('restricted speed', nowadays called 'drive on sight', which was 30 km/h at that time), but this was only realised in the early 1990s. Interviews with local traffic controllers and issuers of rules at that time, revealed that the main reason to do so was 'harmonisation of rules'. The flashing yellow aspect was in the 1990s the only operating mode that required a driver to run 30 km/h at maximum. There were several other operating modes in which drivers should take care, amongst which written orders for passing a signal at danger and for cautionary running, e.g. in case of children playing along the track or similar reasons. Both written orders had a maximum of 40 km/h, yet it remains unsure where this maximum came from. Research revealed that in 1964, the year preceding the change of the meaning of the yellow aspect, the written order for passing a red signal could either be issued for 'restricted speed' (i.e. 30 km/h) or for a speed to be filled in by the traffic controller, similarly to current practices where it is in some cases allowed to run at maximum line speed after the signal was passed with a written order (article 36 Regeling Spoorverkeer). The order for cautionary running did not prescribe any speed, but only a reason for the speed reduction (N.V. Nederlandsche Spoorwegen 1964). Therefore, although no watertight explanation can be given, it cannot be excluded that the maximum of 40 km/h was gradually introduced on the written orders after the change in 1965, so that after 30 years, the notion of 'restricted speed' has been raised integrally, but without proper underpinning, from 30 to 40 km/h.

Education and route knowledge
Interviews with drivers indicated that the current education no longer explains the underlying principles of the signalling system. Instead, it is thought to be sufficient if the driver knows the formal meaning of the signal aspects. This reduces the probability that drivers run with a speed lower than 40 km/h where this is appropriate. A second issue comes with route knowledge. By increasing the speeds such that the visibility distances are too short to react, drivers must already start to brake in anticipation of a signal ahead, which they can only do with route knowledge. Although the intention of the new signalling system was to get less dependent on route knowledge, this knowledge has now become essential in order to stop in front of red signals if these are placed at critical points. When this was questioned in parliament in 2007, this was also formally answered by the railway inspectorate (thereby referring to the former meaning of the yellow aspect that was abandoned some years before; reference IVW-TER/KAB/07/M10.008.102) (Tweede Kamer der Staten-Generaal 2007):

> How is it possible to stop a train, with a braking distance of 400 meters, in front of a signal, when the signal is only visible from a distance of 200 meters*?*
> *The yellow signal aspect means that a driver should decrease speed to 40 km/h, or to so much less as to be able to stop in front of the first signal with a stop aspect. Different than in road traffic, train drivers have route knowledge. They know where signals are placed, and they also know the braking characteristics of their trains. In your example situation this means that a driver knows that he is approaching a signal that may show a stop aspect and from which the aspect can be read from 200 meters. After passing the preceding yellow aspect, he will then continue to brake until the braking distance has been reduced to less than 200 meters.*

However, with the original intention of the yellow aspect, such use of route knowledge was not in question and its accompanying safety risk was therefore also excluded. Later, the adequacy of route knowledge was again questioned in parliament, and the railway inspectorate answered this question by doing research into that. However, the question of whether route knowledge is adequate was immediately rephrased as whether route knowledge fulfils the norms (Inspectie Verkeer en Waterstaat 2009a), so that the question of whether the norms themselves are correct and sufficient was not answered. The report issued by the inspectorate did not indicate what the function of route knowledge is, so that it is difficult to answer the question whether this function is then sufficiently fulfilled, as was the original question. Instead, the inspection concluded that '*there are no specific requirements to achieve, keep and administer route knowledge'*. In that case, it cannot really come as a surprise that no violations of rules could be found, but this cannot be an answer to the question as to whether route knowledge is sufficient to ensure safe operations.

In conclusion, the inspectorate appears to consider the railway system as it exists now as a fault-free 'given' and 'heals' signal visibility problems with route knowledge, instead of uncovering and dealing with such inadequacies at a more fundamental and consistent level as one could expect from an inspectorate.

Increasingly complex station layouts with shorter sighting distances
Some signals are not visible from the required distance of 200 m. After an accident in 1999, where a train SPADed a signal that was visible from only 120 m, the Dutch transport accident investigation board recommended in 2001 that the infrastructure design process should be done in conformance to the safety management system as required by Railned, at that time the independent watchdog for rail safety. Such a safety management system was supposed to be similar to the requirements in the well-known ISO9000-norm (*Regeling veiligheidsattest hoofdspoorwegen* 2004). Also, the accident investigation board indicated that existing designs should be brought into accordance with legal requirements (Raad voor de Transportveiligheid 2001).

Whether or not in response to this recommendation, ProRail's signal sighting requirements have contained since 2003 an appendix that formally allows designs to deviate from the requirements.

APPENDIX A (NORMATIVE) DEVIATIONS FROM THE SIGHTING REQUIREMENTS

(…) The aim of this appendix is to provide a basis for deviations from the signal sighting requirements in order to guarantee sufficient visibility (…).

The requirements for the sighting distance and the continuous visibility cannot always be met.
EXAMPLES: The sight towards a signal can be obstructed by:
- *overhead line equipment in the inside of a curve;*
- *buildings, constructions, screens or walls;*
- *the frequent parking of rolling stock.*

(…) Positioning of a signal in a switching area or at less than 200 m after points can be necessary.
EXAMPLES: Positioning at less than 200 m can be necessary:
- *to achieve short shunting movements;*

> • *to achieve short headways.*
>
> *(…) Point of departure is that the requirements that are not met are fulfilled in a reasonable way.*
> *EXAMPLES: In the following situations the requirements are reasonably met:*
> • *a signal sighting distance of 150 m instead of 200 m;*
> • *positioning of a signal at 170 m after points, while 200 m is required.*

As in the ATB discussion in the 1960s, we see that technical arguments or utilisation needs appear to have prevalence over safety issues. Interviews revealed that these exceptional designs have to be approved by one of ProRail's managers (so-called procedure 00256). That means that the responsibilities are clear, but the question is how this procedure helps to produce an understandable situation for drivers. How can a manager decide about the user friendliness of a design, if the subject matter experts already indicate that they cannot find a solution?

Even though there is such a deviation appendix, it is still not clear why a 150 m visibility distance is 'reasonable' and how this helps the driver to cope with the missing part. Would 120 or 130 m then be reasonable, too? The situation can be safe, though, when drivers run slower than 40 km/h, so as to increase the available reaction time. However, drivers are not explicitly instructed about the cases in which they should do so. There is no special signal aspect that indicates a safe approach speed for these situations. One would expect that those spots where visibility norms had to be bypassed should receive special attention when route knowledge is learned. However, no formal information flow from infrastructure designer to train operating company exists. This means that train operating companies are forced to discover these difficult spots each time all by themselves.

In conclusion, by reducing the required sighting distances the reaction time for the drivers is even further decreased.

Automatic route setting
A final consequence of all this, is that the automatic route setting system does not make a difference between normal signals and signals meant for shunting only (as was shown in Chapter 8) so that the driver will encounter red aspects at difficult positions more frequently.

Conclusion
We have seen that the signalling system 1946 was clear with regard to available braking distance but not with regard to speeds, whilst this was the other way around in the 1954 system. The driver, however, needs information on both dimensions to perform his task. Indeed, modern signalling systems such as ATB-NG, ETCS or the German Linienzug-beeinflussung do provide information about speed and distances; these systems also supervise the braking curve which is also in the speed-distance plane. One might wonder why the necessity to show both types of information was not realised earlier, especially because the need for the re-introduction of a signal aspect 'short braking distance' was already made clear in 1987. Even a concrete plan for introduction was made, but no effective progress has been made since (Middelraad 2000). Another problem appears to be that those who are responsible for operating the current railway system are insufficiently aware of (sometimes implicit) design restrictions, which can than unintentionally be violated.

This example showed that even in the design of safety-critical systems design errors cannot be excluded. Furthermore, they can even remain unnoticed for a long time. In fact, we must

conclude that the original meaning of the yellow aspect was the only one that was correctly underpinned, all newer meanings in fact embroider the error made in 1965. One might wonder why this was not found out earlier. The fault seems to have several causes:

- Implicitly, the system boundaries are changed. Sometimes safety is assessed including the driver's process, sometimes without.
- Insufficient background knowledge of the restrictions of existing systems and their functioning lead to unnoticed violations of such restrictions.
- The drive to utilise the railway to its fullest extent, leads to more and more exceptions being tolerated.

These causes will not be solved within a limited time. Background knowledge will further decrease by retirement of staff, whilst the unbundling of the sector puts emphasis on partial interests and formal agreements, whilst solution of these problems requires an integrated approach and the ability to produce argumentation with regard to the functional contents and not to the formal responsibilities. The governmental wish to further increase network usage will push the system even further towards its limits. Although this is a normal mechanism, a counter-mechanism should be in place, firstly to check whether the boundaries as they are conceived are still sufficient in today's situation and secondly to make sure that these boundaries are not exceeded.

Although these changes are small per se, many of such small changes still lead to a considerable difference with regard to the original intention. Remember that not only the formal meaning of the yellow aspect changed with time, it is now also shown more frequently and in more situations due to changes in the traffic management processes, as was shown in Chapter 8, which may result in drivers being not really alarmed when passing yet another yellow aspect. In fact drivers are becoming conditioned to react improperly to yellow aspects in some cases, while at the same time this signal provides less clarity than it used to. Starbuck and Milliken (1988) describe a similar process of small incremental changes, in their paper on a Space Shuttle explosion, entitled 'fine-tuning the odds until something breaks': engineers and managers feel more and more confident about the system they have made and the successful results they achieve with it and it appears to function well, so they start to cut safety margins. Although these are each small changes, they can quite add up to important issues over time. One of the dangers of this is that at some moment or another the system may be operated beyond its (sometimes implicit) boundaries. Cutting margins is fine as long as this margin is never used, but when the system is safe just because these margins prevent it from being worse, cutting margins is dangerous and should be done only if one is sure that these margins are indeed not used. No such data about how margins are used (e.g. about driver behaviour) is currently analysed on a regular basis, although it is constantly measured for legal reasons.

By increasing the maximum speed after a yellow aspect, drivers can in some cases encounter a lack of sufficient reaction time or a lack of available braking distance. Because of the confusing use of the yellow aspect, which can also mean 'short braking distance', it is also of importance that the driver is educated about what a yellow aspect means in practice, and not only what behaviour the law prescribes. He should regulate the train's speed such that he could stop within the distance over which he can oversee his route. In fact, the conclusion is justified that the introduction of the automatic train protection system has led to a safety flaw in the signalling system, which manifests itself precisely in those situations where the train protection system provides no backup. Also, one could add nuance to the statement that 'no SPADs would occur if drivers were to precisely follow the rules'. Instead, many SPADs are probably prevented just because drivers do not follow the rules too closely and have sufficient insight into the functioning of the system, so that they can adopt a safe driving style in most

cases, even when the signalling system has some hiccups. It is unfair then to find that drivers are required not to violate the rules, while the systems and rules they have to work with are themselves also not free from errors. As a pity, the relationship between signal sighting distance and braking performance normally receives no attention in Dutch large-scale SPAD analyses.

Again, this case indicated that decisions in a safety-critical system were made without fully understanding how the driver, as a user of the system, will be influenced by the proposed changes. The same phenomenon can also be found in the remaining cases of this section, all of which describe how old technology was replaced by newer technology, which inadvertently led to functional changes in the driver's process.

9.4.5 Changing symbol indicator displays to emit yellow instead of white light

Intention
A small, recent change to the signalling system provides an example of a safety analysis which was not carried out from a driver's viewpoint. When Light Emitting Diodes (LEDs) became sufficiently bright and reliable, the idea arose to use them in railway signals. The normal signal aspects (green, yellow, red) were already implemented in LED-technology, but subsequently ProRail also wanted to use LEDs in symbol indicator displays, which are used for instance as repeater signals or to add numbers to the yellow or green aspect (see section 7.3).

Practical use
By legislation, these symbols need to be displayed in white light (*Regeling Spoorverkeer* 2004), following the colour convention that was intended by the designers of the Dutch signalling system, where yellow light is the exclusive colour to indicate braking and white light should be used for additional information (AV130; ProRail 2005a). However, since white LEDs were considered not to be sufficiently reliable, it was proposed to use yellow LEDs instead, based on the idea that the light bulbs that were used up to that moment also emitted yellowish light. To check whether changing the colour would not introduce safety hazards, a safety analysis was made by ProRail (Saarloos 2008), supported by an ergonomics consultant from Intergo (De Bruijn 2007). At that moment 300 units were already installed. However, the safety analysis mainly focussed on the question of what would happen if a driver mistakenly thought that the yellow light would be a yellow signal aspect – in which case he would (unnecessarily) apply the brakes, so that no safety risks were found. Three other questions were not addressed:

1. *Whether a yellow number and a yellow aspect would not fuse.*
   If a signal aspect is often fused with the display indicator aspect when seen from a distance, drivers may be confirmed in their expectation that the distant haze consists of a yellow aspect and a number. This may be dangerous if in exceptional cases the number is not lit (thus a more restrictive aspect is shown). This mechanism (yet without use of LED technology) caused repeated SPADs near Weesp, as was described in section 7.3.2. As was shown in section 7.4.2, the human mind tries to match perceived information to an expected pattern. To counter such a tendency in former times, when a double yellow aspect existed, the yellow lamps were placed as far as possible apart to avoid the different lamps fusing when seen from a distance (N.V. Nederlandsche Spoorwegen 1968).
2. *Whether a yellow aspect would remain sufficiently exclusive.*
   The signalling system is designed such that the colour yellow is exclusively used for instructions to reduce speed, whilst information shown in white provides additional information. Introduction of additional information being also shown in yellow seems to

be undesirable, as it means that drivers have to ignore yellow information in some cases, thus making the 1:1 relationship between 'yellow' and the need to brake less strong.

3. *The interaction with the dynamics of the driving process.*
   Drivers may see a yellow light in the distance and assume that their signal has cleared, and thus they may release their brakes[17], where in fact the signal has not cleared and the yellow light is not the next signal, but an additional message. For example, ProRail's safety analysis report mentioned that if a driver were to interpret a yellow LED repeater signal as a yellow signal aspect, this would reassure him that he has to stop at the next signal, which he already knew, so that no safety effect will result. However, since repeater signals are placed at those locations where the next signal cannot be seen, a safety risk may result if the driver thinks that the yellow light is emitted by the next signal (which in fact still shows red, but is still invisible). He might then think that the signal has cleared and once the red signal becomes visible, he has insufficient time to react (therefore, the repeater signal was placed). Indeed, this mechanism was, more or less coincidentally, reported by a driver who was once misled when he saw yellow light in a bend in a tunnel and thought that his signal ahead was cleared, and then found out that the light was emitted by a repeater signal. Remarkably, one of ProRail's signalling engineers joined the driver who was misled to observe the situation himself and he changed his initial thoughts when he saw what was built. This clearly illustrates the need to share each other's perspectives and to design from a driver's viewpoint.

Lessons learned
We see that the safety analysis was done as a static consideration of stand-alone signal aspects, so that failure mechanisms that involve the real time dynamics (e.g. signal clearances) or interaction with the driver's cognitive processes were not identified. Even though experts from ProRail, ergonomic consultancy firm Intergo and the railway inspectorate were involved in this change, apparently nobody asked the question whether a yellow signal aspect is sufficiently exclusive to detect and diagnose for drivers when other yellow lights are added; the question was mainly whether it fulfilled the design rules and the law. The inspectorate even suggested changing the law to prescribe yellow instead of white light, so that the system would remain uniform (which in itself is of course praiseworthy, but in this case does not solve the driver's possible confusion). None of these institutions apparently was able to oversee the consequences of the change in sufficient detail.

The matter has now been solved since white LEDs are available and ProRail has now decided to use these as quickly as possible.

9.4.6 Losing implicit functionality of automatic block signals

Intention
Running lines between stations are generally not controlled by a traffic controller; they can be compared to a tube. Trains are inserted into one end and at some moment they will pop out at the other end in the same order. The block signals along the line have as (only) function to separate following trains to prevent head-to-tail collisions. When such a line from A to D is obstructed between stations C and D, it is not possible to establish a shuttle service between A and C, although this track section is still free. The reason is that the safety logic of the line only allows one to reverse the running direction of the line when it is completely empty. Also, when a train is stuck on the line and must be towed away by another train coming from the

[17] A braking procedure must formally not be terminated when seeing another yellow aspect, because the next yellow aspect gives no information about the remaining braking distance that is available. However, drivers may know these distances from their route knowledge.

other direction, this train must approach the defect train with written order against the set direction. This was considered impractical and unsafe and hence the idea was proposed to replace automatic block signals by signals under control of the traffic controller. ProRail's responsible safety manager for this project says in an article: '*operating trains on the basis of a fail safe designed system is always safer than operating on the basis of written order'* (Inspectie Verkeer en Waterstaat 2007a).

Practical use
The renewed Amsterdam – Utrecht line, which was opened in 2007, is therefore completely controlled from the traffic control centre, so that for each consecutive block a route must be set, which is normally done automatically by ARI. Now, routes can be set from and to all available blocks. Although this required a total redesign of the safety logic, drivers will not see much difference on the new line, according to ProRail's safety manager. Although the desired advantages have been achieved, other problems came into existence. The former block signals used to be marked with a 'P' (abbreviation of 'permission'), which allowed drivers to pass the signal on sight in the case of signalling failures. This was possible because the running direction in the next block was the same, so no train could come running down the line in the opposite direction. Since this has now become possible (it was even one of the design objectives), every train must now be given a written order for every signal in the case of signalling failures. Drivers complain that although this solution does prevent one train having to run on a written order in the case of a broken train, it requires several written orders for every train in the case of signalling failures on the line. This increases work load for both drivers and traffic controllers.

Some drivers complained that removing the 'P'-marker board on the signal posts removes one of their orientation aids; in other words, the difference in block signals (with 'P') and station signals (without 'P') was part of their route knowledge. If the driver came to a stop in front of a signal without a 'P', he knew that he was standing right in front of the next station, which was especially useful at night. It also enabled drivers to know what to expect behind the signal; if they had to pass it with a written order they knew that they could expect points after passing a signal without P-marking and that there were no points behind signals with the P-indication.

A traffic control related problem is that the automatic route setting system may now set routes in such a way that two trains meet each other nose to nose. Albeit this is not especially dangerous because it is covered by normal signal aspect sequences, it may lead to service disruptions because one of the trains has to be reversed. Another traffic control effect is that the new line is completely overseen only by the Amsterdam traffic control centre, so that the traffic controller in Utrecht now has a poor overview of which trains are actually approaching and how far away they are.

Lessons learned
In summary, although the design intended to better facilitate certain operating procedures, it worsened others, because the improvements were regarded from an infrastructure functionality viewpoint and not from the viewpoint of how to control and use that piece of equipment. Unexpected surprises are the result.

9.4.7 Losing implicit functionality by introduction of GSM-R

Intention
Telerail, the former analogue radio system for communication between traffic controllers and drivers, was at the end of its economic life and had to be replaced by a new system, which

took place between 2004 and 2006. The new system is GSM-R, the new European standard system for these purposes (see section 2.3.7).

Practical use
With TeleRail, there was generally one channel for each traffic controller. Signs were placed along the track at those places where the TeleRail channel had to be changed (automatically or manually). When TeleRail was taken out of service, the signs indicating a channel change were also removed. However, from the viewpoint of the driver, these channel transition signs also (implicitly) indicated that they were entering another traffic controller's area. All this would not be bad, if GSM-R would automatically connect the drivers to the right traffic controller, but this is not always the case, as several interviews revealed. This is especially a problem if several tracks in each others vicinity are controlled by different controllers, for instance on many major stations. Such connection problems were also found in an incident investigation report, although it was not addressed as a shortcoming (Inspectie Verkeer en Waterstaat 2008b). Combined with the fact that traffic controllers do not always mention the area they are responsible for, it may happen that a driver is talking to the wrong traffic controller for quite some time. One driver indicated that *'it is very important in the case of alarm calls that I will be connected to the correct traffic controller and that I can also see who I am connected to. It is not unusual that traffic controllers pick up the phone using their first name, not the area they are responsible for'*. In some cases, GSM-R connects a driver to the traffic controller of the station they have just left instead of the traffic controller who is responsible for the station they are heading for, thus making it impossible to call in an anticipative way.

Also, Telerail offered the opportunity for drivers to hear what was going on, although they could only hear what the traffic controller said and not what other drivers responded. But this often provided them with enough information, for instance about a track defect at a certain location, so that these drivers did not need to call the traffic controller. Such 'listening in' is important to maintain a shared situation awareness, as was shown in other industries (Woods and Hollnagel 2006) as well as railway operation in the United States (Roth et al. 2006). Such shared awareness leads to the emergence of informal correction loops between operators. Although GSM-R caters for similar functionality using group and/or area calls, these functions appear not to be used at the moment.

Lessons learned
Although the speech quality improved, introduction of GSM-R also created some fundamental problems which could even contribute to unsafe situations if a traffic controller and driver need to get in contact with each other and they cannot. The implicit information carried by the channel transition signs, and the need for 'smart' transition points was apparently not identified. Although the new system theoretically provides more functions than the old one, these are not used, so that in effect the replacement led to an effective step backwards considered from an operator's viewpoint. This was, however, apparently not seen from a 'newer is better' technological viewpoint.

In most of the above cases, it was shown that a too simple belief about the driver's and the traffic controller's processes exists, so that the designs or design changes meant to support their tasks lead to surprising and in some cases even counterproductive results.

9.4.8 Discussion on design influences: shift from goal to means

In the preceding sections many examples have been given where the functionality (be it an explicit or an implicit goal) is lost sight of and replaced by the means to achieve that goal.

Thus, the means become an end in themselves. Some projects start with a high level of ambition, but when, during the course of the project, the complexity is too high, the originally defined goal is often narrowed or changes are made such that the original goal is not reached in the end.

Other projects start from a viewpoint of 'replacing old technology' so that the new technology is merely a direct replacement of the old functionality, and new ways of enhancing the functionality by means of the new technology are not always considered. Also, the informal (but valuable) functionality of the old system can by inadvertence be removed because it is not formally known or acknowledged, so that in practice the new system performs worse than the old one. A number of the examples cited in the earlier sections are summarised in Table 9.5. It appears that it is more difficult to maintain a clear picture of a sometimes abstract functionality, as we see for instance with the forgotten function of distant signals (see Chapter 7), the ATB cancel function or train paths; it is easier to imagine a signal as an object, the cancel function as a button or a train path as part of a timetable. Thus, there is a tendency to downscale innovations to simplified concepts that more easily fit current practices – and in the end not so much is improved as was originally planned.

If old technology is to be replaced, this bears the risk that functionality is not conceived from a functional viewpoint of what is desired, but from a technical viewpoint of how it is operationalised and what in that does not work. It is (often implicitly) reasoned that newer

Table 9.5: Intended and realised functionality of several projects.

Project	Intended functionality	Actual realisation
VPT	Fewer communication, better prediction	More communication (due to unbundling), no better prediction (no CS/BO)
BB21	More efficient processes	New technology
2007 timetable	Fixed corridors, higher frequencies, higher reliability	No corridors, hardly increased frequencies, longer travel times, more reliability only after improvements by traffic control (=closing feedback loop)
ATB cancel function	Alert driver on red signal	Push-button
GSM-R	Keep existing functionality	New technology, but with removed trackside driver aids and missed opportunities that GSM-R offers
Infrastructure maintenance	Keep existing functionality	New technology, but without functional improvements for better train paths (e.g. better points to avoid unnecessarily slow driving
Block signal removal	Prevent running on written orders in the case of blocked tracks	More running on written orders in the case of signalling failures; removed orientation for drivers
ATB-EG and changed meaning of yellow aspect	Decrease probability of SPAD	Decreased probability of SPAD at high speed, but increased probability of SPAD at low speed
ATB-VV	Prevent SPADs at highest-risk location	Equipment installed (also) on low-risk-locations

technology must be better, or that whatever fits the rules, must function well. Yet this is not true, as several cases have shown. The pitfall here is that this in fact is based on reversed logic: since new technology has fewer restrictions, it has the potential to serve more functions than older technology has, but it is not guaranteed that these functions are really served by new technology. Similarly, what does not fit the rules has a high likelihood of not functioning correctly, but the reverse is not always valid. Feedback on actual use of existing designs and analysis of the effects of intended changes would appear to be a worthwhile improvement.

9.5 Design and management in safe envelope terms

Designers must provide tools and procedures that help the processes during operations. So, a first prerequisite is to know these processes well. The examples in the previous sections have shown that this is insufficiently the case. Taking the theoretical framework into account, designers have to make sure that operators can see the safe envelope's information items (e.g. the envelopes, the position of the process therein and how this situation will change). Table 9.6 provides an overview of the conclusions on the safe envelope requirements for the train driving process, the rescheduling process and the route setting process, as these were identified in Chapters 7 and 8.

In principle, the entries in the table that are marked with 'no' need to be improved by better designs or procedures. Some of the projects mentioned in this chapter had the aim to do so, such as the BB21 project or the CS/BO-module in the VPT system; however, these projects were not brought to a successful conclusion. Current projects of which some try to improve the situation are mentioned in Appendix A, such as RouteLint or the European Train Control system, which respectively try to display more information about the current state and about the operational envelope to the driver. Many other projects described earlier in this chapter are hard to map on this table. Often, this is caused because they were based on a technical change, so that improvements of the items in the table were not taken as a deliberate starting point.

Instead, the railways spent much emphasis on making sure that the driving process would not exceed the operational envelope in the spatial/physical domain, i.e. that a train run is not performed beyond the limits of the movement authority, as the developments of automatic train protection systems show. However, several changes in the past decades have gradually led to a change in the operational envelope which allows more system states than what used to be allowed, in some case even without clear consideration of whether this would remain controllable. In certain circumstances insufficient control information is provided to keep the process within these new envelope's boundaries (e.g. changed meaning and more frequent use of the yellow aspect; deactivation of the ATB cancel function).

Also, it is a task of the railway's management to make sure that the functions in the cascade model cooperate smoothly. There is hardly any effective process control that keeps trains away from safety-critical boundaries and on their desired paths (causes: poor information; poor predictive tools). The process thus unnecessarily often approaches the operational envelope's boundaries, marked by yellow and red aspects.

We have also seen that there is only limited feedback from the operational level to the design of support systems (e.g. automatic block signals, GSM-R), which could serve to make such problems recognised and addressed by a redesign.

Table 9.6: Overview of safe envelope analysis for support of three operational processes.

Information item provided	Rescheduling (time/spatial domain)			Route setting (spatial/physical domain)		
	Driver	Local traffic controller	Regional traffic controller	Driver	Local traffic controller	Regional traffic controller
Actual state	Driver knows where he is, but he can only see whether he is on his path at discrete points for which a timetable entry exists	No explicit insight into the time dimension and no conflict detection	No; only implicitly (plan + devia-tions), no exact track use or conflicts in the plan	Available at discrete points (lineside signals)	Yes (routes set + train locations)	No; no informatio n on track use.
Goal state	Available at discrete points for which a timetable entry exists, but not up to date	No (plan becomes goal)		N/A	Yes (plan lines in ARI)	N/A
Operational envelope	No, the plan contains no explicit tolerances	No, the plan contains no explicit tolerances		Yes, movement authority given by main signals	No; cannot see which alternative routes are also acceptable	No
Attainability envelope	Not provided (e.g. amount of running time margin)	No, the plan contains no explicit information on buffers and margins and conflicts arising during operations are not indicated		Yes, implicitly provided by signalling distances which allow for attainable braking distances	No, there is no updated real time feasible schedule	
Controllability envelope	Based on driver experience (e.g. slopes in track, slack in timetable and braking characteristics of train)	No information provided on last moment to solve a conflict (e.g. before train reaches distant signal)		Last moment of intervention indicated by distant signalling and by driver experience	No information provided on last moment to set a route (e.g. before train reaches distant signal)	
Viable envelope	N/A	N/A	N/A	No need for this when the train is supposed to remain within the operational envelope.; see also Table 7.6	No, but only re-quired in emergen-cies; otherwise n/a	N/A

9.6 Underlying causes: management responsibilities

In this section, the findings of the previous three chapters will be summarised under the headings of twelve underlying causes. These are displayed in Figure 9.4 and form, for the moment, the final stage of the grounded theory approach. This means that the available data from the interviews and literature did not yield any further explanations or further abstractions during my analysis. This does not mean that more could not possibly be found if the same material were to be reconsidered from a different point of view or in relation to other data that were not used during this research. Table 9.7 provides an overview of how these underlying causes can be traced back to the information provided so far; each of the causes will be explained in more detail below, where the section numbers refer to the number in Figure 9.4.

9.6.1 Goals not clear

On several levels in the railway industry the goals that must be achieved appear not to be clear. On an organisational level, NS appears to struggle with the question whether trains should be punctual or fast, let alone including how this relates to the passengers in these trains

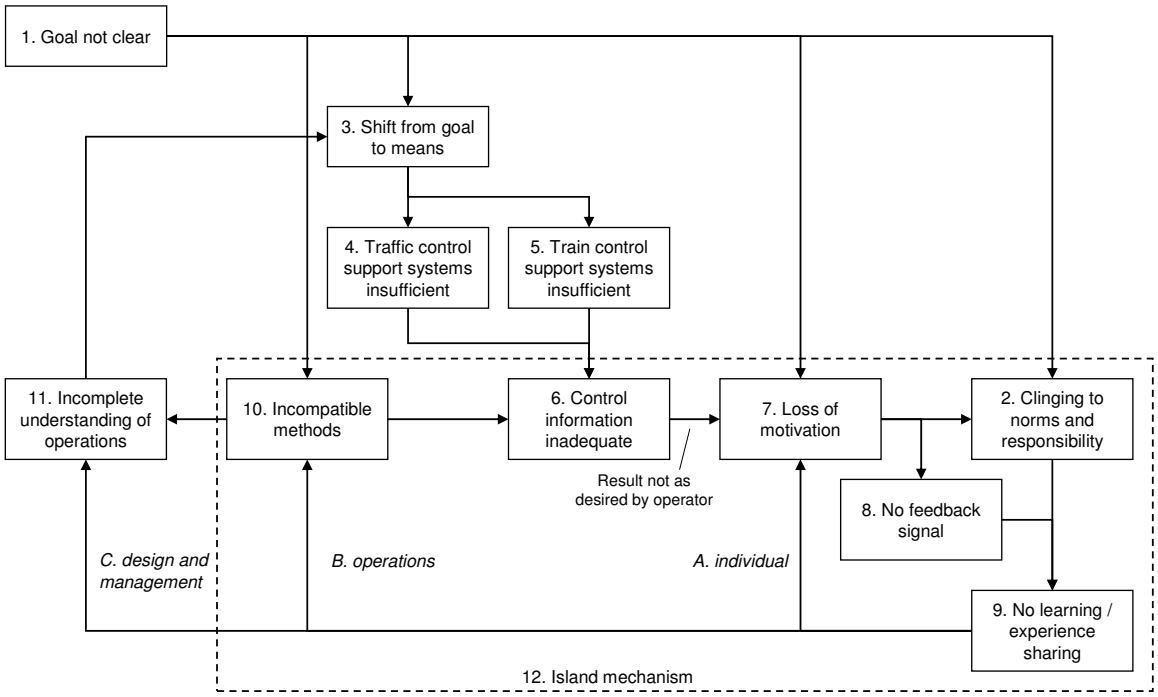


Figure 9.4: Cause-effect diagram for underlying causes of SPAD and quality problems.

Table 9.7: Overview of cases related to the twelve underlying causes in Figure 9.4.

Underlying cause:	Reason for inclusion (described earlier in):
1. Goal not clear	Organisational level: top level KPIs still changing (2.2.5), lack of vision (9.2.1); Infrastructure design: timetable and operational processes unknown (9.3.1); Timetable design: shortest possible running times without deviations (9.3.3); Regional traffic control: focus on freight trains (8.3); Traffic controller: set routes or solve conflicts? (8.2, 8.3); Driver: regulate timing or speed? (Chapter 7); Operations: plan becomes goal (7.2.1, 8.3.2).
2. Clinging to norms and responsibility	Conservative attitude (9.2.1); Infrastructure designers: stick to design norms (9.3.1); Timetable designers:: stick to planning norms (9.3.3); Local traffic controllers: stick to plan and TADs (8.3.2, 8.3.5).

3. Shift from goal to means	See Table 9.5 on page 205.
4. Traffic control support systems inadequate	Unpredictable behaviour of automatic route setting system (8.2.6); Poor oversight leads to few plan modifications (8.3.2, 8.3.3); Communication times increased, leading to slow decision-making (8.5.3); Conflicts not shown in actual plan, as was originally intented (9.3.4); Advanced support as intended by the BB21 project has been cancelled (9.3.2).
5. Train control support systems inadequate	Do not show path deviation and rescheduled path (Chapter 7); Advanced support as intended by the BB21 project has been cancelled (9.3.2); Difficulties to define most relevant speed profile (7.3.5), due to ambiguity in yellow signal aspect (7.3.3) and relationship between route and signal (7.3.4); On approach of a red aspect: mismatch between allowed speed, braking distance and sighting distance (9.4.4); ATB cancel function removed instead of being made more effective (9.4.2); Automatic route setting system leads to build-up of driver expectancies (7.4.2, 8.2.5); Over-reliance on route knowledge (7.4.4).
6. Control information inadequate	Logical result of the preceding two issues and the incompatible methods during operations.
7. Loss of motivation	Drivers and traffic controllers are disappointed about the fact that they do not get sufficiently informed about relevant changes in cooperation with traffic controllers; both are disappointed about suggestions for improvement that are not being picked up (section 8.5).
8. No feedback signal	Logical result of preceding issue; see also section 8.5.
9. No learning / experience sharing	Learning requires that operational surprises are reported and analysed (section 5.4.5). The presence of many different behaviours under traffic controllers (8.5) and incompatible ideas between traffic controllers and drivers indicates that there is no sharing of experiences. This point was also mentioned by many interviewees directly.
10. Incompatible methods	Actors have different ways to cope with timetable deviations, which makes co-operation difficult. Timetable designers do not make the slack time available to operations (9.3.3); regional traffic controllers act once the delay has grown above a threshold of 5 minutes, yet they only do so for freight trains, whilst local traffic controllers are confined by predefined dispatching rules, which, however, are not implemented in the automatic route setting system (8.3); Train drivers try to reduce their delays, whilst traffic controllers and the automatic route setting system take them for granted (8.3.3). Thus, conflicts may be solved on paper, but may remain in practice (7.2.2). Also, many practitioners developed different ways to cope with problems (8.5.4)
11. Incomplete understanding of operations	The lack of organisational learning provides no information on how the system actually functions, which leads to incomplete understanding of the operational processes. Several examples have been given where this leads to unforeseen results, e.g. in timetabling (7.2.2, 9.2.2, 9.3.3), in understanding the reason for setting of partial routes and their difficulties (8.2.7), or in the design of several safety systems, including the ATB cancel function, the changed meaning of the yellow signal aspect, or the deployment of ATB-VV (9.4).
12. Island behaviour	Island behaviour is the self-reinforcing process described in section 8.5 and consists of several of the aforementioned underlying causes.

and their goals. Besides that, another goal is that NS should now be profitable, so that making money also becomes a goal. ProRail used to consider itself as an institution that was meant to manage an infrastructure network instead of as a company that delivers services to customers, for which the infrastructure network merely functions as an important means of production. This now appears to be in the process of change. These different organisational goals can conflict and, since the unbundling, there is no longer a central board of directors to cut the Gordian knots of conflicts, especially those between different, now independent system actors such as infrastructure providers and train operating companies.

Also on the level of individual actors, such as drivers and traffic controllers, there is uncertainty about which goal they have to pursue, as we have seen in Chapters 7 and 8. Is it a driver's task to regulate the train's speed, or the train's adherence to its path? Is it a traffic controller's task to set routes, or to solve conflicting infrastructure claims? As a result, actors formulate their own (sometimes incompatible) goals: for instance, timetable designers want the shortest possible running times, regional traffic controllers try to find solutions for heavy (freight) trains but do not focus on individual passenger trains; local traffic controllers focus on their own area (some of them minimising delays, others minimising the number of changes to the plan) whereas train drivers focus on their train trip through a number of areas.

9.6.2 Clinging to norms and responsibility

Because goals are not clear, operators might decide to cling to their formal responsibilities, so that they cannot be blamed for not doing their work properly. Also, traffic controllers indicated that it is possible to hide your own failure in the organisation in this way. It is always possible to blame someone else. Clinging to formal responsibilities is also visible on the organisational level, where ProRail and NS are quite good at focussing on the KPIs which they have agreed with the Ministry of Transport, yet all other possible goals on which there is no KPI (e.g. 'short travel times' or (formerly) 'good train paths') will then not be strived for, although the latter now is being changed.

Several examples have been given of undesired results in practice that were fed back towards designers, for instance regarding brittle timetable design or poor signal visibility, yet often the answer is that 'the situation satisfies the norms' thus giving the signal that the reported problem cannot exist because it is already known that the solution must be good. When this sort of feedback does not lead to a change, this again discourages those who initially reported the undesired situation. Reports of such situations ask for a reconsideration of the norms rather than that these norms can be used to justify whatever goes wrong in practice. Thus, satisfying the norms also becomes an end in itself instead of a means to lead to proper results. In terms of organisational learning, this is called double loop learning.

9.6.3 Shift from goal to means

As was summarised in section 9.4.8 many examples can be given of projects where the original intention was replaced by the surrogate need to bring the project to an end.

9.6.4 Traffic control support systems do not fit the task

Traffic controllers do not have information about all dimensions in which they have to work, being the longitudinal and lateral track layout, and the time spans during which track sections are reserved for train movements or for other purposes. Regional and local traffic controllers each miss one of these dimensions so that they are dependent on each other for a complete overview of the situation and possible consequences of decisions they make. This leads to long communication times or, if such communication is skipped, to a plan in which conflicts remain. The presence of conflicts in the plan is not explicitly indicated to controllers, even

though it is their main task to solve these. Another complicating factor for local traffic controllers is that they are sometimes faced with unpredictable actions of the automatic route setting system.

9.6.5 Train control support systems do not fit the task

Chapter 7 showed that drivers also do not have sufficient information to fulfil their task properly, which requires information on whether the actual timing, position and speed match their requirements. Drivers lack information about the adherence to their (re)scheduled path and about where slack time is available to cope with delays or to give the opportunity to drive slower than the maximum speed in order to prevent unplanned stops.

With regard to the signalling system, it was found that the choice for a speed signalling system created uncertainty about available braking distances: thus, the yellow aspect can indicate many different situations. The red aspect provides information on the required stopping position, but it is not always visible from the required distance or it cannot be unequivocally related to the track the driver is running on. This requires drivers to have route knowledge, even though it was one of the intentions of the signalling system currently in use to be less dependent on route knowledge.

9.6.6 Control information is inadequate

Network and traffic controllers as well as train drivers only have limited situation awareness, i.e. having a perception of the actual situation and a prediction of the system behaviour if a certain option is chosen. This was also found in earlier research in the Netherlands and in other countries (Lenior 1993; Banverket 2002; Biemans et al. 2005; Kauppi et al. 2006). Prediction of the future is mostly based on the experience of the staff with earlier similar situations; sometimes such experience is required because actual information is not available, and operators are forced to operate on the basis of 'default values'. If the (mental) representation of the situation is not correct, it is also impossible to make sound decisions on how to control the process, so that from time to time undesired results occur.

9.6.7 Loss of motivation

For several reasons, operational results can be different from what the operator wanted. It may occur that the operator has formulated his own goal and sees that the result is not as intended; it can also happen that formal regulations forbid certain control measures in certain situations (e.g. waiting for a connection or deviating form TAD procedures in the case of delays), or that the desired result can only be achieved with excessive effort, which is clearly also not desirable. Although most interviewees do not mention it literally, the overall message is that operators are inherently motivated to improve the system, but are in fact losing control over the system. A SPAD is one of the safety related symptoms of this uncontrollability. Other symptoms, such as unnecessary waiting times or waste of traction energy by unplanned stops are symptoms from a quality of service viewpoint.

The operators' inherent motivation may suddenly change into discouragement if they have put effort into improvements, yet they experience that this did not help. Some operators then restrict themselves to their formal responsibilities and simply let the undesired results happen again and again: 'they can get it as they want it'. Some interviewees indicated that the railway industry is not critical enough: *"Others easily say: 'let it be, it's OK'. Whoever wants to perform better than the average is easily perceived as a troublemaker"* and improvements are thus curbed instead of promoted.

9.6.8 Hardly any feedback

Loss of motivation leads operators no longer to provide feedback, since they feel that this is useless. Usually there is no feedback from one actor to another. For instance, traffic controllers do not know whether train drivers are satisfied with their solutions; nor do the timetable designers know about the controllability of the schedule they have made. Also, there is no regular exchange of information between train drivers and designers of track layout and signal positions, which was felt as a serious shortcoming, not only by drivers but also by one of the most experienced infrastructure designers. Almost all drivers and many traffic controllers mentioned that they used to report undesired results, often accompanied with suggestions for improvement, but that their suggestions were hardly ever realised and that they were not even rewarded for taking the trouble to report the situation. This appears to be especially the case if the current answer to the reported situation is not changed because 'it satisfies the norm', as was already indicated at the corresponding sections above; thus indicating that the receiver of the feedback is in fact not open to learning and change.

9.6.9 Learning and experience sharing amongst actors

The railway system is controlled top-down, which requires the top level actors to have a relatively abstract model of operations. Thus, it is inevitable that discrepancies between intended operations and actual operations result. Feedback from operations (or any other actor) is required to update this model and decrease the amount of discrepancies; this is a normal learning process. If there is no feedback, the discrepancies remain intact. However, such feedback is often not present, so that there is no organisational learning and no continuous improvement, which is more or less the same since 'improvement' means to know how to do things smarter than before. In particular double loop learning does not occur (questioning and adjustment of goals and resources; see section 9.6.2.)

In this chapter, several relatively important changes on various functional levels in the rail sector have been described, which all aimed to improve performance. There appears to be a tendency to start each time from scratch (e.g. BB21, timetable 2007), yet this bears the risk that no analysis is made of why current practices do not lead to the desired result, and thus operational problems are insufficiently understood. Often such projects fall back into old practices and/or the focus shifts from improving performance to developing new technology (e.g. BB21, ATB-VV). With regard to traffic control processes, such as the behaviour of the automatic route setting system or the paradoxes mentioned in Chapter 8, it would also be useful to perform such analysis in order to understand and improve these processes. In each of these cases, the railways insufficiently go back to the question 'what is going on here?'. Often there is no systematic problem analysis, but instead they take a shortcut to their preferred solution. Although that appears to be a pragmatic and efficient method on first sight, it does not lead to better understanding and thus not to the ability to come up with smarter and more effective solutions. An example is the fact that the SPAD workgroup quickly decided to develop ATB-VV, supposing that this would be the most effective measure. However, no underpinning for this belief could be found and, indeed, later analysis showed that this solution will not be as effective as was widely believed (Inspectie Verkeer en Waterstaat 2008h). The workgroup did not start to look for SPAD mechanisms such as those listed in Table 7.4 in which the dynamics from the driver's viewpoint play a role, but only did so at the moment that other measures apparently were not leading to the intended improvements. The analysis, done at that (late) stage is a good example that shows that such an integrated approach can be fruitful.

Learning leads to increased knowledge. With regard to current knowledge of the system several peculiarities cannot remain unmentioned:

- Only a very little knowledge about how and why technical designs work as they do is written down. This is true for the Dutch situation, as examples with regard to the changed meaning of the yellow aspect or the ATB cancel function have shown, where changes have led to safety hazards; but the point was also raised by Pachl (2008) as a general concern in the railways of other countries. From this study, the fact that several professionals did not realise the fundamental functional difference between a distant signal and a main signal is worrying, since this is the key concept of all signalling systems in railways.
- Many employees in railways have gradually made themselves familiar with the system's functioning by (internal) education or by working their way during their careers through many functions starting as driver, traffic controller or maintenance engineer. This provides excellent, though largely implicit situation awareness, but also bears two risks: many older employees are already retired or will retire within a few years, taking their knowledge with them. The other risk is that the knowledge of many employees has developed based on certain designs and design choices, so that these are often regarded as fixed or that the consequences of changes cannot be overseen since there is no model for such a situation.
- Knowledge about why the system is as it is, seems to disappear gradually, as we have seen in the examples of route knowledge, which is assigned an unintentionally high importance to 'justify' why signals need not be visible from their required distance. Yet this is not how the signalling system was initially intended; also the knowledge that not all routes that are accepted by the interlocking can be used for train movements is fading away. Disappearing knowledge is also mentioned as a problem in a recent analysis that was carried out for the railway inspectorate (van Bruxvoort et al. 2009) and was already addressed in parliament questions in 2003 (Schultz van Haegen 4 July 2003).
- It appears that several problems that exist today were solved in earlier days; in other words, the railways then forgot why they had made certain choices and adopted certain solutions and unknowingly took them away again. Examples are the signal aspect 'short braking distance', the fact that it was not always allowed to rely on the interlocking only, and the fact that it was under certain circumstances required to inform the driver when route setting would be different (i.e. shorter) than he would expect and when he could not see this from the signal aspects.
- Even the railway inspectorate, whose task it is to check whether the companies under their supervision act in a safe way, appears not to learn. When the inspectorate answered parliament questions about the deactivated ATB cancel function, or about the insufficient signal sighting distances in relation to the required braking distances, the answers given appear to defend the system's weaknesses rather than try to solve any of the mentioned problems.

In summary, the conclusion is justified that the railways do not learn easily from their errors.

It is remarkable that the railway's approach to solving problems is not very scientific. This is demonstrated through the use of blurred definitions (e.g. for partial routes, vigilance, route knowledge, capacity) which cause confusion. Another indication for a not too scientific approach is that a proper problem analysis is often lacking, so that there is incomplete understanding of why things do not lead to the expected results. Thus, often a shortcut is taken from a problem to a possible solution. Such a shortcut bears the risk that such solutions are not well thought-out or make other things worse, as various examples in this chapter have shown. Improved assessment and management of design changes is recommended. Taking a more scientific approach certainly does not require difficult formulae, yet it would be a good starting point to ask first 'why' a problem exists and only afterwards 'how' it can be solved and which unintended side-effects a possible solution may have.

9.6.10 Incompatible methods during operation

The fact that there are different, unclear goals and that operational surprises are often not addressed and improved, results in several incompatible methods during operations; these again lead to inadequate control information for other actors because incompatible methods make the system's behaviour hard to predict and because information is not available in similar terms. As an example, several actors assess whether a train adheres to its planned path, yet all of them have different ways to cope with deviations from the intended path:

- Timetable designers only cater for deviations of up to a minute or so by providing slack time. However, the amount of slack time added is not explicitly clear during operations;
- Regional traffic controllers will assign a new path once the delay has grown above a threshold of 5 minutes, yet they only do so for freight trains. Slack time cannot be explicitly taken into account; instead, delayed trains are quarantined;
- Local traffic controllers are confined by predefined dispatching rules, which, however, are not implemented in the automatic route setting system;
- The automatic route setting system sticks to the planned order and routes of trains; and does not take slack time into account;
- Train drivers react to delays at certain points and try to minimise them using slack time, yet they do not precisely know where slack is built into their scheduled path. This is incompatible with traffic controllers who take delays for granted; these approaches can impede each other.

Regarding safety, most traffic controllers rely on the interlocking in normal operations – a technology that ensures that routes can only be set in accordance with the applicable procedures. Thus, they believe, no invalid signal aspects can be shown and as train drivers must obey the signals, safety must be assured. However, some traffic controllers and about half of the train drivers emphasise that it is important that traffic controllers realise how their actions change certain conditions in the train drivers' process. Additional information may help drivers to anticipate situations. One case was reported in which traffic controllers rely on train drivers obeying their signals, whereas train drivers can only obey their signals correctly under the assumption that the traffic controller did not set a wrong route – but that is not checked by the interlocking. Thus, in this case traffic controllers and drivers are mutually relying on the other one to make no error.

Many drivers are not satisfied at all with the solutions provided by ProRail traffic control. Drivers said that traffic controllers lack creativity and do not think ahead, which causes unnecessary delays. Also, nearly all drivers and traffic controllers mentioned that they wanted to have more knowledge of each other's work and some of them have organised this when they were off duty; sometimes even although this is forbidden (e.g. for a traffic controller to join the driver in his cab). This clearly indicates the need for sharing experiences.

9.6.11 Incomplete understanding of operational processes

Due to the fact that there is limited feedback from operations and because there are incompatible methods during operations, it can be hard for managers and designers to understand actual operations. Since it is not always clear for designers and managers how operators behave, this behaviour is often covered by the blanket term 'expertise', which in fact demonstrates that operator experience is considered as a black box. Notwithstanding the usefulness of skilled operators, this reliance on expertise as black box leads to a situation in which operators are insufficiently supported in their tasks, because those who have to supply support cannot fully understand how to do that. An example from this research is the high value that is attributed to route knowledge, even though nobody could clearly say what this exactly is or why it is needed. Also a British study about the nature of expertise in railways

concluded that *'Network Rail must take a very serious and hard look at what its expert operators of the future should be and how to help them develop'* (Farrington-Darby and Wilson 2006).

Several findings of this research suggest that on the design and management level there is only a limited understanding of actual operations and that goals are not always specified in a functional way; for instance it is surprising that drivers and traffic controllers do not have supportive tools for their main tasks, which are to regulate the trains' deviations from the scheduled paths and to avoid conflicting infrastructure claims. Certainly, tangible and imaginable concepts like 'speed' or 'route' are easier to understand than 'relative timing' or 'path conflict', so that it requires a somewhat more abstract view to get these concepts into focus. In contrast to the fuzzy operations, norms and guidelines provide a clear grip on what to do and what not to do. Yet the risk with such norms is that, because they are also made by designers and managers, they still do not provide enough support to cope with the daily dynamics in the system. An example is the control square or the predefined dispatching rules for traffic controllers. Also the common explanation for partial routes and the recent discovery that the ATB-VV patch will probably not be as effective as everybody hoped, can be considered as examples in which the mental model of designers appears to fall short: apparently, the model that formed the base for that hope is an inaccurate one, so that feedback from operations and a more thorough design management process will be necessary to find such inconsistencies at an earlier stage.

Several changes to safety-critical systems, such as the changed meaning of the yellow aspect, the introduction of the yellow LED symbol indicator displays or the removal of the ATB cancel function are poorly underpinned. The solutions appear to be more or less decided upon up front, and the underpinning is something that is tailored to the solution afterwards. This does not mean that the system is made unsafe with intent, yet it shows that designers and/or managers who decide on such topics apparently have a strong conviction of their own correctness and lack a sufficiently scientific approach to analyse the problem before they come to a solution.

9.6.12 Island mechanism: actors cut themselves off to keep oversight

Finally, as a combination of several factors, operators tend to cut themselves off from each other; in Chapter 8, we called this 'island behaviour'. This is caused by several factors outlined in earlier parts of section and forms a self-reinforcing phenomenon, as Figure 9.4 shows. When operators feel discouraged and stick to their formal responsibilities and do not interact with other operators, they also force other operators onto islands, making the system more complex and harder to control.

9.7 Conclusion

This chapter has showed how design and management cannot be seen apart from operational problems. In summary, the following main causes for quality and safety problems were identified in the previous three chapters, using a grounded theory approach, and linked to each other (see Figure 9.5):

1. Goal not clear
2. Clinging to norms and responsibility
3. Shift from goal to means
4. Traffic control support systems inadequate
5. Train control support systems inadequate

6. Control information inadequate
7. Loss of motivation
8. No feedback signal
9. No learning / experience sharing
10. Incompatible methods
11. Incomplete understanding of operations
12. Island behaviour

As can be seen from Figure 9.5, three vicious circles were identified which keep problems intact, and should therefore be broken, for example by organisational learning:

A. Individual loop (loss of motivation);
B. Between operators (incompatible methods);
C. Between operators and designers (incomplete understanding of operations and lack of predictive or corrective design reviews).

The rail sector tends to look for add-on solutions: how to make sure that a driver does not pass a stop signal in a given situation? However, we need also to look for front-end solutions: how to prevent a driver from running into such a troublesome situation of having to stop unplanned at a signal? Both ways need to be combined. The latter approach requires that operational envelopes which are modified by several processes are geared to each other. Currently, this is done implicitly in the minds of staff and hence the precision and effectiveness of the traffic control process depends to a great extent on the experience of staff working in it. It is unfortunate then that this staff is insufficiently supported in their tasks and many are discouraged to suggest improvements and retreat to their islands.

Since disturbances cannot be taken away and spread themselves quickly in highly used networks, they must be explicitly taken into account – planned for, made controllable and managed. However, current infrastructure and timetable designs are made independently of the traffic control philosophy and without taking possibilities for efficient traffic and train control tasks into account. Several safety checks and principles that are used during timetable

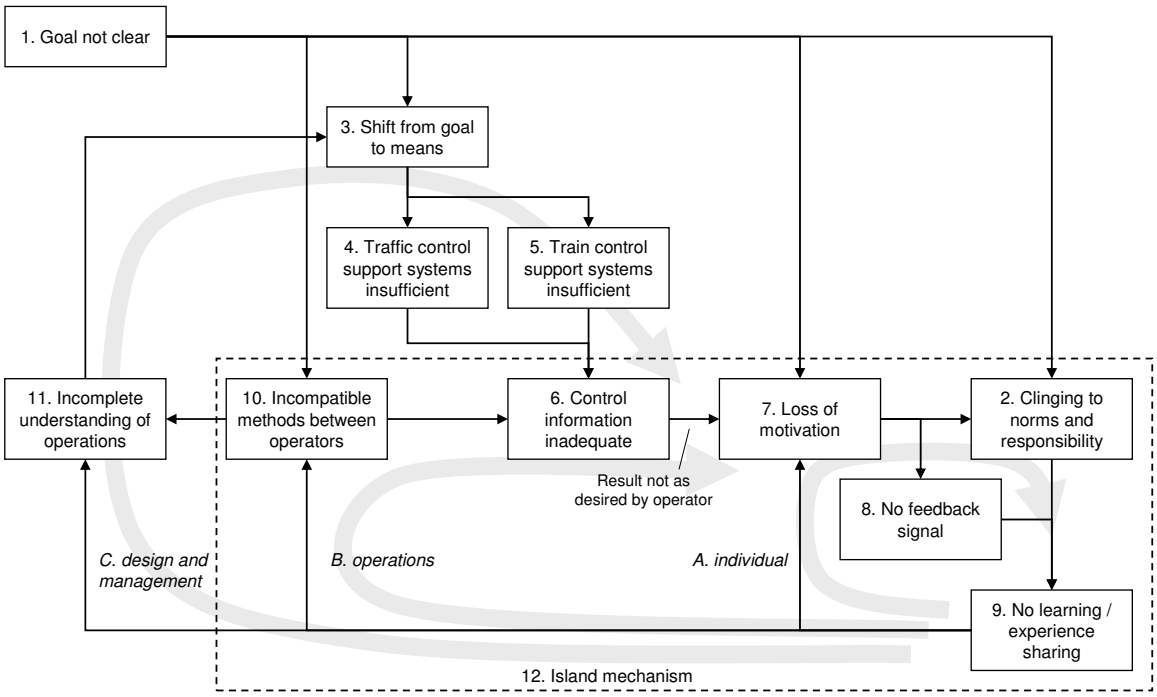


Figure 9.5: Twelve underlying causes with three vicious circles indicated.

design are in fact correct, yet they would make more sense if these were (also) performed during operations. However, even though the traffic controller's main function now is to solve conflicts, intended aids to make these conflicts visible and show consequences of possible solutions have not been introduced. Thus, no updated, conflict-free, real time schedule is made when disruptions render the old schedule infeasible. This again makes the traffic process difficult to predict.

Even though it is a main task of the driver to run according to schedule, he is not informed about changed schedules, nor can he see in sufficient detail whether he does adhere and how he can adhere better to the scheduled path. This results in an unnecessarily high number of restrictive (yellow or red) signal approaches; also the fact that trains do not behave according to the plan makes the traffic process more difficult to predict for traffic controllers. Since timetable designers and traffic controllers all rely on the driver to obey signals, a correct detection and interpretation of signals by the driver is still a single point of failure. However, imperfections in the route setting system, the signalling system and the train protection system, and the impossibility of faultless operator behaviour make this a dangerous assumption:

- The yellow signal aspect is not clear in all cases. Causes are the changed meaning of the signal aspect itself, as well as its increased and sometimes even incorrect use.
- The identification of the red signal aspect is not reliable under all circumstances. The approach speed and the sighting distance are not compatible, so that a driver's route knowledge has unintentionally become an essential element in order to be able to stop in front of some red signals.
- Train protection systems such as ATB are often favoured by designers to solve safety problems; yet such systems do not directly help to fill the identified gaps in process control and sighting distance; in the best case they make sure that these gaps do not result in accidents.

A main problem is that designers cannot sufficiently understand operational processes, since there is hardly any feedback to them about it, whilst the island mechanism is self-reinforcing and therefore needs to be broken in order to incrementally provide more motivation, more compatible methods, and thus feedback for learning. The next chapter will provide some concrete 'handles' and examples on how to improve.

PART IV: Discussion and future

10 How to improve safety and quality of service?

10.1 Introduction

In the previous chapter, twelve underlying causes for SPAD and quality problems were identified. In this last part of the dissertation, the findings of the previous chapters will be glued together. The knowledge from the theoretical chapters and knowledge from the practical chapters on current railway operations will be combined, so that it will become clear how process control can be improved to increase safety and quality.

The chapter is structured as follows:
- Section 10.2 provides a short summary of the findings of Part III of this thesis.
- Section 10.3 provides a reflection on the framework used and suggests some modifications to it.
- To make sure that the functions in the cascade model (Figure 2.3) can work smoothly together, it is necessary to provide a clear goal, to motivate staff to do what they have to do, and to actively learn from problems during these processes. These can be considered a prerequisite for the ability to improve the other underlying causes that were shown in Figure 9.4. This chapter will therefore start with suggested changes to the organisation as a whole (section 10.4), which is primarily a management responsibility.
- Section 10.5 focuses on coherent flows of information between processes and functions in the cascade model.
- Requirements from the theoretical framework will be used to formulate improvements for traffic controllers and drivers in sections 10.7 and 10.8. This basically comes down to supplying operators with information about where the process they have to manage is, relative to the actual goal and relative to the positions of the boundaries stemming from earlier processes. Several existing proposals to enhance these processes will be briefly mentioned and placed into this context.
- Section 10.9 will discuss several issues with regard to how technological developments do, could or should support operators.
- The chapter concludes with a summary of the main findings of this chapter.

Several railway employees indicated, in reviewing the drafts of this thesis, that managers would be interested in costs and benefits of proposed solutions. Such information is not the

field of this scientific research but a normal business decision. Appendix E provides some information on this topic, and results from several publications which show that better control can reduce delay minutes by 60%, reduce SPADs by 50% and lead to energy savings up to 30%.

10.2 Short summary of current operational control problems

In the previous chapters, several problems in the control functions were found. The plan (timetable) becomes the goal for traffic controllers, since they do not know the requirements that were the basis for the timetable design. Therefore, they also cannot see whether they will harm the service intention, for instance if they break a passenger connection. Not only can Dutch traffic controllers hardly see whether the system goal is met, it is also difficult for them to see conflicts before they become acute, and it is also difficult to see the consequences of possible solutions. As a result, they often make as few changes to the original plan as possible, and conflicts may remain.

Drivers are not informed in detail about how to stick to the conflict-free path that should ideally result from the scheduling and rescheduling processes. The driver thus has no information on how to navigate through the operational envelope: being late or being before schedule can be considered as the driving process running outside the operational envelope stemming from traffic management functions. In other words, the traffic management function(s) assume certain behaviour of the trains they try to control, but they do not provide sufficient information to make sure that this behaviour will indeed result. Although most traffic controllers think that drivers are fully informed by the signalling system, this is not the case. Under such a regime, the cooperation between traffic control centre and drivers is bound to be imprecise. Current practices can only work if there are few disturbances, which is increasingly untrue as network usage intensifies.

Although the deviations may appear small, in the order of magnitude of 1 or 2 minutes, this may be enough to occupy the next train's path and the disturbance will spread, and the process will move unnecessarily often towards the operational envelope boundary stemming from infrastructure allocation functions (e.g. yellow and red signals). Buffers, in time and space, are currently used to reduce the probability of such effects being passed to other trains, but the deviation from the desired path and the fact that capacity remains unused in buffers is not prevented by doing so. This disturbance can be amplified if a train is required to brake and needs a long time to re-accelerate. Such delay propagations can lead to chaotic system behaviour, in which a small initial disturbance may lead to unpredictable behaviour – not so much for the driver, who is only concerned with one train, but for the traffic controller and the customers. Delay propagations should be considered as an internal process disturbance which can have severe effects on various indicators of rail operations:

- *Quality of service*: repetitive unplanned braking and accelerating leads to discomfort and uncertainty for passengers and as such is a threat to quality perception (see the criteria in section 2.2.2).
- *Safety*: several examples of rail accidents in the Netherlands show that a number of small delays that are all within acceptable margins (e.g. the punctuality threshold), may still lead to uncontrolled interaction with an unforeseeable result, which can be an accident (Van den Top and Groeneweg 2007; Inspectie Verkeer en Waterstaat 2008f).
- *Staff motivation*: a system that behaves chaotically is also unintelligible for the staff working in it. People who have no way of finding out whether the action they are about to perform will make the situation better or worse are hard to motivate to take responsibility

for overall outcomes. Hence their ability to take responsibility strongly correlates with their ability to make informed decisions.

Increased network load increases the probability of trains affecting each other and would require larger buffers – whilst at the same time increased network load means that buffers should and do decrease. Ashby's Law says that the amount of required buffer time is inversely related to the amount of control that is exercised. Because no control mechanism is able to exercise perfect control, there will always remain small, uncontrollable disturbances for which small buffer times should be provided. But still, we can conclude that managing the proper use during operations and thus reducing the amount of buffer times is the key to increasing network load safely. This means that a higher network utilisation is not a problem per se; it only becomes a problem when this leads to uncontrollable situations.

Where control measures fail, the backup functions serve to stop trains that run into a remaining conflict:
- Since today's interlockings work mainly error-free, these will not be further considered in this chapter (see section 8.2.1).
- The Dutch signalling system provides in general clear warnings on running lines, but not always in stations where the situation is more complex. This means that the safety system's operational envelope is not always clearly indicated to the driver. In some cases this is caused by difficulties in reliably detecting the stop signal itself. In other cases information that was provided by the preceding signal(s) with a yellow aspect can be ambiguous.
- The automatic train protection system works well on running lines, too, but in the more complex station areas its information is also limited, so that safety ultimately relies on the driver.

In several cases, exceptions or deviations have been normalised, such as the yellow-yellow-red signal aspect sequences, the limited speed that was raised from 30 to 40 km/h and the fact that routes that are accepted by the interlocking are now considered safe, whilst there were additional restrictions in place in former times. The system has changed; the higher utilisation asks for quicker response times and improved situation awareness, whilst the current situation with many actors involved in the decision-making process leads to long response times and hence the inability to combat disturbances effectively.

10.3 Reflecting on the theoretical framework

10.3.1 Usefulness of the framework

In this work, the concept of the safe envelope of operations (see section 6.2.1) was further developed in order to develop new insights to (and from) the field of safety science. The contribution of this work to the development of the safe envelope can be found in section 4.8. Whilst the safe envelope framework makes clear the boundaries within which the process must be controlled and the main information items that the controllers needs to have to succeed in doing so, the model should be combined with some form of a process control model to indicate *how* the process will be controlled within the boundaries of the safe envelope model.

The fact that DDA sequences can be found in safety science, management science, human factors and engineering is a big step forward, as it allows for integrating insights from these necessarily diverse disciplines into a coherent model capable of working across system levels and combining technology, human behaviour and organisational decision making. The safe

envelope with its DDA loops helps to achieve a broader understanding of the control processes, but the same approach can also be used to analyse for instance management processes.

The cascade model connects processes within the wider framework, whilst the safe envelope allows the in-depth analysis of which information is required to fulfil a function. The combined framework, consisting of safe envelope, DDA sequences and the cascade model was applied to railways for the first time, taking an integrated view: it looks not only at technology or human factors, or at safety or capacity in isolation; but these are all considered together. The approach allows us to find accident causes also found by current methods (section 6.2.2 – concurrent validity). Its explanations, however, go further than what we see currently in the railway sector (e.g. public accident investigation reports by the inspectorate and internal investigations by ProRail or the SPAD workgroup). The explanations also functionally describe how barriers – a common concept in safety science – ought to function and that active and passive technological barriers can also be replaced by active (behavioural or automated) control. Thus, it was possible to make a synthesis that has not been made before: both safety and quality can be seen as outcomes of a control process and occurring problems as failures of this process. Improvements in the control system (which also comprises management) such as closing the feedback loop within operations, but also from operations to design and management, will be necessary to solve the problems.

10.3.2 Application of the safe envelope

Application of the safe envelope framework proved to be useful to convey the insight that both safety and production issues relate to process control; whilst safety mostly benefits from avoiding undesired states, quality of service benefits from reaching desired states – both of which can be taken into account by controlling the system properly. Although the principles of the safe envelope were clear to most people, some indicated that they found the pictures hard to interpret. A possible improvement to the framework could be to enhance the ability to display cross-sections of the safe envelope. However, also without drawing the pictures of envelopes, the framework could be used to identify gaps and problems in the current control structure of the railway system, as is for example shown in Table 9.6 on page 207, by outlining whether several controllers do have information on the information items that follow from the envelope framework.

Several processes, such as maintenance, shunting or using level crossings have been excluded from analysis in this thesis (see the introduction), but it was suggested that the same approach will work with them. Indeed, the same approach does appear to be possible. Since damage can be considered as unintended transmission of energy between two objects (Johnson 1973), the objects and energies available in (or in the environment of) the system can be used to see which types of energy must be controlled and which objects are at risk. The other way around, it is also possible to derive from the available types of energy which type of accidents may occur. For instance, to avoid derailments, the energy transfer (force) between rail and wheel should not exceed a certain value, which requires control of axle loads and centrifugal forces. If we want to prevent collisions, we need to make sure that no two track users have access to the same piece of track simultaneously (hence it is impossible to exchange energy between them, whether they are other trains, road users or a maintenance gang occupying a length of track). We also need to ensure that every user does not leave the safe zones (e.g. track section(s)) assigned to him. In generic terms, these requirements are not different for maintenance gangs or level crossing users than for train movements; however, the functions and the restrictions under which these functions must be executed may be different so that other safety measures may result. Further analysis along the lines of this thesis is needed to establish this.

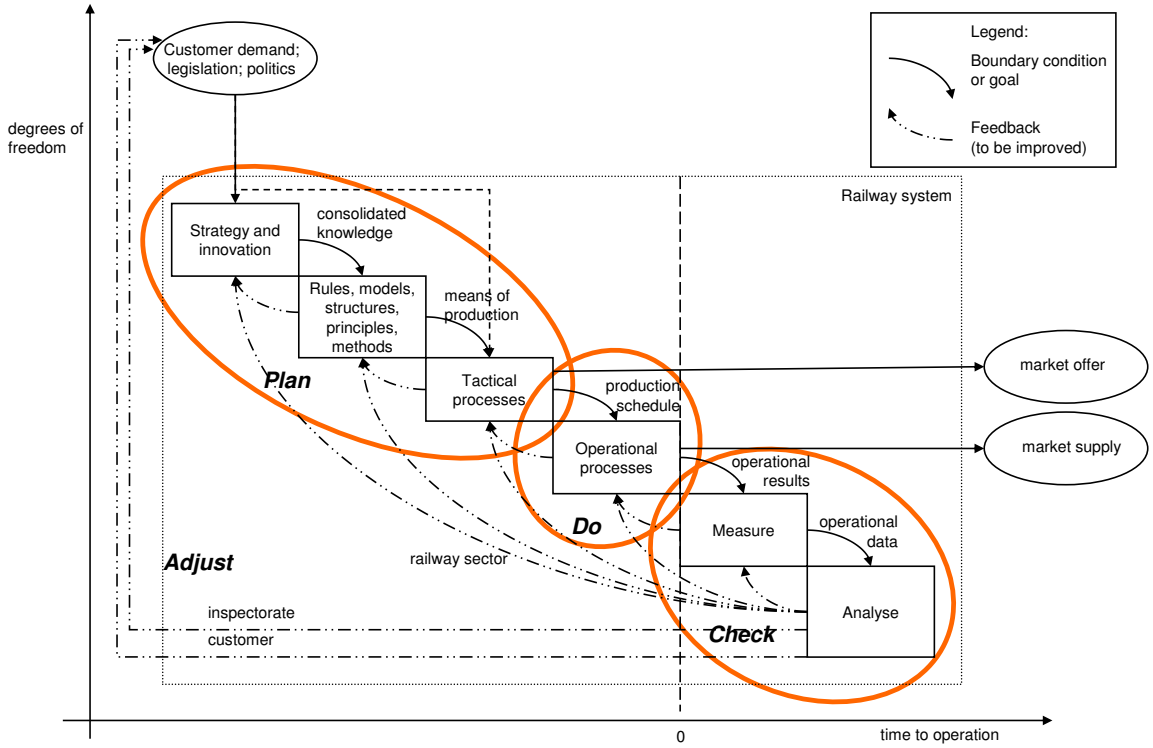


Figure 10.1: Extended cascade model.

10.3.3 Further developments to the cascade model

Based on the grounded theory and discussions with colleagues, amongst whom Arco Sierts made a major contribution, the cascade model can be further extended. As introduced in section 2.3.1 it represents the IST-situation, but after modifications and extensions it can also reflect the SOLL-situation, which is more practicable for suggesting improvements. The result of this extension is displayed in Figure 10.1.

The functions that were used in the model thus far in Figure 2.4 are now grouped under the 'tactical processes' and 'operational processes', where tactical process intend to plan how resources will be used and operations has the task to execute the plan and cope with disturbances, which may require adjustments to the plans (Schaafsma 2001). These are now part of a greater diagram, which as a whole is also a plan-do-check-adjust loop and constitutes the whole railway system. From the top of the diagram, requirements on the railway system are posed, from customers and from the society. The Plan part of the system consists of three processes:

- Strategy and innovation: this functions aims to find smart ways to do whatever is required, by proposing adjustments to methods and means of production such that these are better able to lead to the services that must be delivered.
- From this preceding process, models, rules, structures, principles and methods are derived. These summarise the knowledge from the previous function in a more concrete form. These models and rules indicate which means of production are required and which services will be offered. Examples include the Rail 21 future plan, which asked for four-track sections and level-free junctions in order to operate a three-train-type system. Its successor, Benutten en Bouwen, relies on the principle that no new infrastructure shall be built and aims to operate a highly reliable two-train-type system (e.g. only local and intercity services). Rules, structures, models and methods enable the organisation (but also human beings) to perform actions without the need to analyse a situation each time again

and again. This is efficient, as long as the rules and models are still correct. The processes before tactical processes are normally slow processes and they appear semi-static. This bears the risks that changes in these processes are not considered even when they are necessary. To make sure that the models and rules are still correct, the feedback loop at this level should be closed. All of this implies that the organisation should have in place the two elements of a successful design management process, namely predictive and corrective design reviews, seeking to predict in advance any unwanted effects of proposed design changes and to steer the design process as it is carried out, to ensure that it does not deviate from its planned objectives along the way.

- The tactical processes define which means of production will be assigned to which production processes at what moment; their output is a market offer and a production schedule. So far, everything constitutes a plan. Frequently, the first two functions are bypassed and demands are directly put into the tactical processes. Although this allows for quick reaction to changes in demand, it bears the risk that people do not think deeply enough about the need for or value of more fundamental changes. A filter should be installed that lets only small change demands get through unanalysed at this level.

The next set of processes consists of the operational processes. These correspond to the 'do'-phase and execute the production schedule. Although they are only a tiny bit of this diagram, this is the process where the added value for the customers is really produced. Their outputs are operational results: a market supply.

After the operational processes, the next two functions perform a check: to do so it is necessary to measure results in the first place; then, these must also be analysed. It is not only the railway company itself that should analyse its results; also customers and society do so, albeit sometimes in an implicit manner. As far as safety is concerned, it is the task of the railway inspectorate to measure and analyse on behalf of the society. The need for expanded and improved measurements and analyses has been shown throughout this thesis. Measurements only make sense if it is known what must be measured. Also, (societal) pressure to perform may lead to good news shows, e.g. those variables are measured or presented on which improvements are visible (Gelders et al. 2008).

To close the feedback loop, the results of the analysis are fed back in order to give new insights to all preceding functions, including to design, a loop that is notably absent at present. In that case, organisational learning becomes a fact. Feedback performed by customers and society can also result in other products or services being asked for at the most fundamental and abstract level of the model.

A last remark is that the cascade model may incorrectly give the impression that only one actor can be active on every level, which is not the case. In fact, similar configurations of actors are currently present in a number of railway-related organisations (e.g. also in train operating companies) so that there in fact is more communication between actors than shown in the pictures used in this representation may suggest. Although this representation proved useful to give an explanation for current performance problems, a more detailed picture may arise by further investigation of information flows. Examples are naturalistic or ethnographic studies in the traffic controllers environment (Farrington-Darby et al. 2006; Farrington-Darby and Wilson 2009), or cognitive work analysis (Vicente 1999).

10.3.4 Usefulness of the grounded theory approach

With hindsight, grounded theory was a useful method to combine highly detailed data from several sources into a relatively manageable number of recurring factors: the twelve

underlying factors that contribute to quality and safety problems. These twelve factors must all be turned into something positive; it is insufficient to only pick out the 'most important' ones, since the factors are intimately connected. Not addressing one of the causal factors will cause other factors to persist as well. Especially, three vicious circles were identified which must be broken to improve performance (see Figure 9.5):

A. Individual loop (loss of motivation);
B. Loop between operators (incompatible methods);
C. Loop between operators and designers (incomplete understanding of operations and lack of predictive and corrective design review).

As can be seen from Figure 9.5, these three loops have in common that there is hardly any feedback and organisational learning. Since all loops are extensions on loop A, a first prerequisite is that operational staff must be motivated to report such surprises, and that such reports must be analysed and fed back to other operators and to system designers, but also, vitally, to the reporters to indicate that their ideas are being considered and have been used. This allows for gradual optimisation of the processes in the cascade model (section 2.3.1), such that not only will every actor attempt to optimise his own box, but that every actor contributes to optimising the overall result. If there is insufficient understanding of actual practices, this leads to a situation in which no useful and targeted improvements can be introduced. Therefore, learning from operations and maintaining knowledge about system design and its underlying design assumptions is of utmost importance. It is also important to realise that quality and safety (and their potential absence) are in fact created at the end of the chain of tactical and operational processes. All other processes are merely a preparation to enable traffic controllers and drivers to perform in an optimal way.

These issues will be addressed in the next sections, where the working hypothesis is that each section needs the preceding problems to be solved in order for its own issues to be resolved and to function correctly.

10.4 Unravelling the problem (A): Encourage operators to report surprises

The first loop of self-reinforcing causes and effects is that of the individual operator. Many operators lost their motivation and therefore do no longer supply feedback. But without feedback, the situation cannot be improved, which again leads to operators losing their motivation. There appear to be two ways out of this vicious circle, which are both a management responsibility. The individual loop (A) remains intact because operators are discouraged to report problems. As can be seen from Figure 9.5, one of the causes for this is that there is no clear goal for them to achieve.

10.4.1 A clear goal to optimise the result

Without a clear goal, the activities of an organisation or an operator cannot be directed, nor can the quality of the results be assessed. As we have seen in Chapter 2, railways can be seen as a system that is organised around time: a timetable is central both for customers using transportation services, and for the railways to produce these. The essential variables can be derived from the services as they were advertised in the timetable, and are *(arrival) time* and *position*. These are the main variables of which a trip (in the transport market) and a train path (in the path market) is constituted, so that these variables should be used to define the quality of service. In order to provide services with a predefined quality, it is necessary to prevent these services from varying too much from the intended goal. The acceptable variance around

the service goals can be displayed in an operational envelope; so, ProRail as a supplier of train paths should agree with train operating companies not only on a scheduled path, but also on acceptable tolerances around the path. A possible goal for traffic controllers could then be to keep trains within these agreed tolerances as much as possible.

10.4.2 Staff motivation

In order to achieve the goal and avoid staff retreating to their 'safe' but restricted islands, it is necessary to have motivated staff. Ellenbroek (2005), referring to Merchant and Van der Stede (2003), indicates that management can perform three types of control measures to achieve the system goals, which have different effects in staff motivation:

- Action controls, which have a direct effect, but require a detailed understanding of the processes. Also, since they prescribe what to do, action controls may have a negative effect on staff motivation and creativity.
- Result controls, which do not require precise understanding of the activity and are widely applicable. They allow for creativity of staff which can motivate them. A difficulty is that these controls cannot be used, as long as the result is not clearly formulated; also, it must be possible for the operator to affect the result, and the result must be measurable.
- Personnel and cultural controls are meant to enforce that the desired behaviour of workers is achieved 'automatically' through a set of common ideals. Merchant and Van der Stede (2003) indicate that an organisation should always start with the question whether such personnel and cultural controls would suffice to achieve the goal, because they are cheap and have hardly any negative side-effects.

In the previous chapters, we have seen that traffic controllers are mostly supposed to apply standard scenarios in the case of train delays, which is an action control. In order to increase their motivation, it seems to be worthwhile trying one of the other types of controls. A concrete suggestion will be made in section 10.5.1. Another source of discouragement comes from poor cooperation between actors, which makes it difficult for any single operator to achieve the result as desired. Therefore, breaking the second loop (B) by enhancing the cooperation between actors will also be helpful. This will be described next.

10.5 Unravelling the problem (B): Improve cooperation between operators

In Chapter 8 it was described that operators have developed highly varying routines, and the conclusion was drawn that there is a need for less ambiguous practices but without throwing the informal working methods, despite their resulting ambiguity, overboard: enforcing a model could break fruitful routines. The puzzle is not to find the optimal control model directly, but to make sure that the organisation learns how to develop and preserve fruitful routines. Through constant learning, a better model may come within reach incrementally. Two types of information flows appear to be essential:

1. *from operator to operator (which is loop B)*: organise joint learning that includes all operators concerned. For example, simulate scenarios with service disruptions in a team of different operators. When recognising informal routines, the pay-off should be checked from multiple perspectives.
2. *from operators towards designers and managers (which is loop C)*: enable operators to contribute to process innovations. If bottom-up feedback from operations does not reach the innovation processes, misconceptions between designers and operators may harm either the functional workarounds in operations or the implementation of a promising innovation, as Von Meier (1999) and Busby and Hibberd (2002) describe. To avoid such discrepancies and their resulting operational problems in the first place, new or modified

designs should therefore be subject to a predictive design analysis, for example a hazard and operability study (HAZOP), as is common in the chemical industry. A modified version of this procedure, the Traffic HAZOP, has been developed by Jagtman (2004) for road traffic infrastructure design.

10.5.1 Control functions throughout the railway's production process

For drivers and traffic controllers to know what is the right action to be taken requires them to have good situation awareness, as was shown in Chapter 5. The safe envelope of operations can help us to derive which information is essential for operators (drivers or traffic controllers) to know. To achieve their goals, they must know the goal and its allowed tolerances, which we called operational envelopes. To remain safe, they must be able to know where the viable envelope is and at which moment or place action ultimately must be taken in order to keep the system in a safe state. Perceiving and diagnosing this information in turn requires correctly functioning DDA loops (detect – diagnose – act), and since much information stems from other functions, these DDA loops should be coherent throughout the cascade model. We can conclude from the previous three chapters that the railways try to ensure safe operations by performing three DDA loops, which they consider to be independent of each other. These are:

1. timetable design; in order to solve conflicting infrastructure claims and grant train movements at their desired time. Traffic control is used to handle exceptions.
2. signalling, to provide the train driver with a 'stop' message in those cases where a train path could not be provided in time;
3. automatic train protection, to stop the train in those cases where the driver does not react adequately to the signalled message.

Although these functions are certainly necessary, they do not cover everything needed. Also, it is clearer to divide the functions into either control functions or backup functions, where the control functions aim to reach the goal within the operational envelope, whilst the backup functions serve to correct control errors and aim to prevent the process leaving the operational envelope. The result is shown in the functions listed in Table 10.1. The functions that are currently not supported are underlined and should be introduced in process control.

Table 10.1: Control and backup functions.

	Control function (aims at goal within op. envelope)	Backup function (guard against leaving operational envelope)
Timetable design	1A) Path definition	2A) Ex-ante conflict estimation
Traffic control	1B) Conflict resolution and rescheduling 1C) Route allocation	2B) Real time conflict estimation 2C) Interlocking
Train driving	1D) Remain on path	2D) Signalling and train protection

The function 1A specifies a 'draft' of intended operations. Since this is a preparatory task, it is useful to check whether the draft provides a sufficiently workable production schedule during operation. To do this, an estimation of the number of conflicts that will arise during operations should be made: this is the new function 2A. Traffic control (1B and 1C) realises the production schedule by turning the draft into a final version (rescheduling, which mainly consists of conflict resolution), and setting routes accordingly. In doing so, backup functions are 2B and 2C, of which the latter must be considered as vital, to avoid errors in one of the preceding functions resulting in unsafe routes. Function 1D is realised by the driver and

backed up by function 2D, which makes sure that the driver does not leave the operational envelope as defined by the route assigned to his train and are also vital.

Each of the functions can be represented as a DDA sequence. As was already mentioned in section 1.1, hardly any cases are known in which inability to act is the cause of SPADs. The failing parts of the DDA loops are primarily related to detecting and diagnosing information; so that the SPAD problem is mainly an information exchange and information prioritisation problem. Note that this holds not only for drivers, but also for traffic controllers and designers. If the above-mentioned control loops simultaneously fail to maintain a safe situation, incidents and accidents could occur. Note that quality of service is already at risk once the control functions do not work properly. In that case, unnecessary conflicts may remain in the plan which can lead to delays, even when the situation can be kept safe if the backup functions are sufficiently reliable. Quality improvements require the control functions to be improved in any case. As a side-effect, fewer exceptions will remain for the backup functions. In that case one should also be aware that the risk of 'out-of-the-loop' problems with the humans involved in the back-up functions may increase, and more use may be necessary of simulation practice to resolve this.

Current technological developments lead to an increasing integration of timetabling, traffic control and train control functions and their backup functions, regardless of whether this is the case on newly-built high-speed lines (Grant 2003), on existing rail networks (Pachl 2005) or on metropolitan networks (Plamann and Klindt 2008). In all cases, the aim is to create integrated information flows between all relevant actors. The following subsection will show how control functions can be fluently connected. Improvements for the operational functions of traffic control and train driving will follow in sections 10.7 and 10.8.

10.5.2 Interacting control processes in railways require congruent information flows

In railways, only approximately 25% of train delays are caused by incidents and defects, the remainder being a mismatch between the timetable and the real process, for which several causes can be found (Weeda and Hofstra 2008). In other words: with their actions some actors in the railway's control processes disturb the control processes performed by other actors. We have seen examples and explanations of this phenomenon in Chapters 7, 8 and 9. This illustrates the need for compatible actions and compatible information flows between the different functions in the cascade model. Figure 10.2 shows an example of how subsequent control processes interact by specifying time windows for the next sub process.

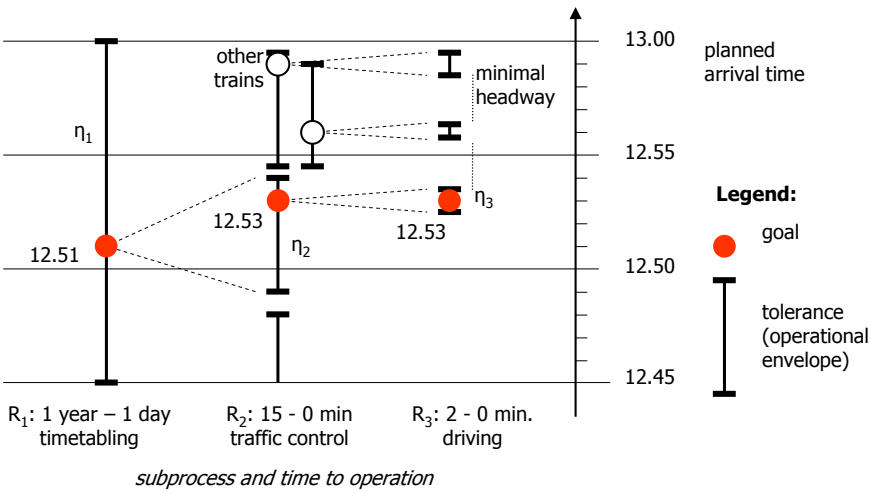


Figure 10.2: The tolerance is narrowed for successive sub-processes.

The function of the *timetabling process* is to create service patterns that fit customer requirements. In this example it is assumed that passengers want to arrive at the latest by 13.00 in a certain station to catch a connecting service. They certainly do not want to arrive later, but also they do not want to wait very long. We therefore assume a 15 minute time window, from 12.45 to 13.00, in which the arriving train must be scheduled in order to be a satisfactory service (η_1 in Figure 10.2). This is what the Swiss railways define as the 'service intention' (Wüst et al. 2008). It serves as a functional requirement specification for timetable design. Due to other constraints, the timetable planner (R1) chooses the published arrival time to be 12.51. Once published to passengers, this time provides a narrower (and currently mostly implicit) time window for the traffic control process (R2 and R3), for example from 12.45 to 12.54 - assuming that a 3 minute delay is still acceptable to hold connections and fulfil passengers' needs. Given the departure times from previous stations, it may not be feasible for the train to arrive earlier than 12.49, which effectively narrows the time window for the traffic control process even more (12.49 – 12.54).

It is the function of the *traffic control process* to regulate the process such that it remains within this new time window (η_2), even when disturbances such as train delays and track failures occur. Due to disturbed traffic, the traffic control centre may reschedule the train to arrive at 12.53, just before trains that are scheduled to arrive at 12.56 and 12.59. The service intention, which serves as a functional requirement specification for timetable design, is currently usually not known during rescheduling; however, it should be kept to because, also during re-scheduling, it would be necessary to see whether the new plan still fits the service intention.

The traffic control process again specifies a narrower (and again currently mostly implicit) time window for the driver (η_3). He should keep the *driving process* within this time window in order to comply with the updated plan (Weits 1997; Schaafsma 2001; Laube et al. 2007). The situation becomes complex if a train fails to stay within its time window. Then, it affects the traffic controller's ability to set routes for other trains in time. An effective control system must therefore attempt to limit such unexpected restrictions to a minimum. This means that the train driver should explicitly know the boundaries of the train's time window, in which he can operate without (unintentionally) disturbing other train runs. In the current situation, this is not the case: he does not know these limits explicitly. In order to connect the train driving process to the traffic control process, it must be easy for the traffic control centre to adjust the schedule and it must also be possible to provide it easily to drivers. A current (but still not fully operational) project in which ProRail is trying gradually to achieve a similar situation is that of 'customer oriented logistics'; see Appendix A1. A comprehensive Swiss study indicates that such an approach can lead to significant improvements in network capacity and reliability (Lüthi 2009).

Time windows can be seen as an agreement between the supplier and the customer of a service, in which is negotiated what is wanted and what is possible (Weits 1997; Schaafsma 2001). They indicate the tolerated variety in output quality that is still acceptable. Time windows are not to be confused with buffer times, as the latter are reasoned from the viewpoint of the production system itself (which is then idle) rather than from what it should achieve. Having buffer times does not indicate that customers will be happy: even if there is plenty of buffer time, quality of service can be poor, if trains are delayed more than what is tolerable from a business viewpoint. Having buffer time makes it likely that delayed trains will not delay other trains, so they allow the making of a statement about the stability of the production system. Another drawback of buffer times is that they make the system slower. It makes sense not (only) to prescribe a certain amount of buffer time to be included in the

timetable, but a certain amount of disturbances that the system must be able to handle without affecting the service quality as defined by the service intention (Wüst et al. 2008).

10.5.3 Gearing operational envelopes

As we have seen, there are several operational envelopes, stemming from different functions in the cascade model and serving to indicate which boundaries are imposed on the operational process. Some stem from limitations in the business goal, such as time windows agreed with train operating companies. Others, such as maximum line and train speed stem from limitations of the means of production, or from how these are assigned to different production tasks (e.g. route setting). All operational envelopes have in common that they are defined by an earlier function in the cascade model and indicate what can be tolerated.

To prevent damage involving train movements, the amount of (kinetic) energy in the system must be controlled (e.g. braking curve supervision). Since the mass of a train is given once it is on the move, the kinetic energy can only be controlled by influencing the train's *speed*. Also, to prevent physical contact between a train and other objects it must be made sure that every piece of infrastructure is assigned to only one user at a time and no user leaves the assigned section(s) of track. Therefore also the *position* must be controlled. Speed and position are therefore essential to control safety. Time and position, and speed as the variable linking these two, were defined earlier as the main variables for service provision.

The variables '*position*' and '*speed*' therefore have two different envelopes within which they should be kept: one for safety and one for quality of service. Safety and quality of service cannot be controlled independently of each other. The envelopes for speed and permissible train position should be (partly) overlapping in order to be able to steer the process in a safe way towards its goal. An example is a train that should depart, since it will otherwise be delayed, but whose signal is still red. There is clearly a conflict between the two envelopes at that moment. A hierarchy must be used for which operational envelope should certainly be met in any given case. Even better is to avoid requirements from different envelopes conflicting. This means that the operational envelopes should be geared to each other. If they are not, the process cannot be executed in such a way that all constraints are met simultaneously. In the case of erroneous control information, the process might be steered towards and ultimately across the viable envelope boundary and damage can result, unless there is a 100% watertight operational envelope that stops the process before damage happens (e.g. a fault-free safety system). As long as this is not the case some kind of information filter is required that lets information only pass if it is not conflicting with operational envelopes that stem from safety-critical restrictions.

10.6 Unravelling the problem (C1): Learning from actual operations

10.6.1 Organising feedback and organisational learning

In different parts of the railway organisation(s), different conceptions exist of how the system works. Not all problems can be seen in all conceptions. This causes some safety risks to remain unsolved or even to be created when decisions are taken to change the design of systems, the timetable, the operational rules or the organisation itself – examples are the different ways in which actors cope with delayed trains, or how changes to the signalling system were made without fully taking the effect on the driver's process into account. It is important to align the perceptions of all actors that affect one process, such that no problems will arise on the operational level. One of the most important ways to do so is to feed back problems experienced at any level to those actors that decided on this topic earlier in the

(design) process, so that designers and operators have similar conceptions about how the system is intended to work and how it actually works.
Breaking the third vicious circle requires that operational processes are better understood so that they can also be supported by better designs or design changes. The ability to do this builds on the achievements of the first and second loop being broken, so that operational processes function more as a whole and problems therein are being reported. As was described in section 5.4.5 about organisational learning, a learning agency should facilitate dealing with 'operational surprises'. An example of a learning agency within ProRail is the performance analysis unit in the traffic control department, yet this unit only focuses on a limited number of operational surprises which are mainly related to timetabling issues. Also the expert groups that look at a detailed level into the causes of SPADs at multi-SPAD signals can be considered a learning agency. It would be useful to extend performance analyses also to other areas, such as signalling and infrastructure layout, and for operations support systems such as the behaviour of traffic control staff and the automatic route setting system. Also a safety management system can be considered as a superimposed control loop, and thus as a feedback mechanism. It is a means to regulate the behaviour of technology and people in the system (Hale 2005). ProRail's safety management system is also based on Deming's Plan-Do-Check-Act control loop (ProRail 2006). However, its effectiveness as basis for organisational learning is doubtful, since little of the data entered into its corresponding database does indeed lead to recommendations for improvement (Beekman 2008).

If there is no effective learning, this can also lead to loss of motivation of staff; either because they do not want to report the same errors again and again, or because they get no reward for taking the effort to report problems. The existing performance analysis unit, the SPAD expert groups and the safety management system are a good start, yet it would be even more fruitful to interconnect and expand these. Chapters 7 - 9 also serve as such a feedback, from which some suggestions for improvement can already be given with regard to improved process control and the solution of incompatible methods.

10.6.2 High Reliability Organization

If ProRail, or the railway sector in general, wants to go a step further than organisational learning, they could aspire to become a High Reliability Organization. These were already briefly mentioned in Chapter 3 as organisations that share certain properties to cope with tightly coupled and complex systems. Weick and Sutcliffe (2001) provide a list of five of such properties (in italics):

1. *Preoccupation with failure,* which means that organisations must actively look for warning signs, analyse them and act on the analysis (this also is a DDA loop).
2. *Reluctance to simplify interpretations,* wich means that HROs try to use all available data to create a more complete picture. Where cost-cutting organisations may regard such elaborate data processing as redundant work, HROs consider it as a vital double check to avoid disaster.
3. *Sensitivity to operations,* which means that the island-mechanism as described in Chapter 8 must be avoided, since it hampers operators in seeing the broader consequences of what they are doing. Also managers should encourage operators to report on their experiences and they should provide enough capacity to analyse operator's signals.
4. *Commitment to resilience*: HROs do not regard errors as undesired events that should not have happened – much more, they know that errors cannot always be prevented. Thus, the organisation must have the ability to quickly recover from them and learn from them about its own (mal)functioning.
5. *Deference to expertise,* which indicates that during calamities the decision-making power must be located with those people who have most expertise to solve the matter. Higher-

level authorities are temporarily by-passed. This is important to be able to counteract disturbances more quickly than they can develop, as was shown by Aulin (see section 5.5.2) – in tightly coupled systems, this therefore needs to be done very close to the point at which they occur.

10.6.3 Define the processes first, then the technical solution

As this thesis has now made clear, the definition and understanding of the functions that should be supported must be the starting point for technological developments. If this is done, technology can support business processes and can be tailored to whatever is required. The opposite approach, starting with the building of 'better technology' or 'enhancing capacity' under the assumption that newer technology will automatically lead to better results, has been shown to lead to undesired results in several examples in Chapter 9.

For example, timetable and infrastructure design should explicitly take into account how the operational processes can deal with delayed trains. Traffic controllers must find new paths for trains, so it must be explicitly considered and ensured during design that they can successfully perform this aspect of their tasks, as far as reasonably practicable. It is necessary to develop dispatching rules *before* the final timetable design is determined. Minimum headways and buffer times, the current criteria for timetable design, provide insufficient guarantees that solutions can be found during operations. If conflicts are likely, the service intention time window (e.g. earliest possible departure time or latest arrival time of a train) could be stretched somewhat. Although this makes the service offer less attractive, it provides more freedom for traffic control to reschedule trains, so that they can also find a solution more quickly.

Improvements in infrastructure design should be aimed at allowing faster cycle times (e.g. headways) so that more rescheduling options can be found within a given time window. Prototypical situations can help several actors to easily understand the situation. For example, dividing the infrastructure network into congested area's (bottlenecks) and repair zones (buffers), as is intended in the Dutch idea for customer-oriented logistics (see Appendix A1), can help infrastructure designers, timetable designers and traffic controllers to choose appropriate measures in each zone. On a somewhat more concrete level, templates could be defined for standard tasks that the infrastructure must support, such as passing a local train on a double track line, a meeting of two trains on a single track line, or reversing a train at its endpoint. In order to allow a flexible use of these measures during operation and to allow for a high utilisation, an aim should be to design such prototypical situations such that their processes can be executed in the shortest possible time.

For regional lines, several such examples have already been developed by Sierts for ProRail's contest on how to make regional lines cheaper and more attractive (Sierts 2007). For main lines, an example is that intercity trains can overtake local trains with hardly any time losses, in contrast to the same manoeuvre currently, which requires approximately 4 minutes in addition to the normal dwell time of the local train (Van den Top and Sierts 2009a). However, saving these 4 minutes requires that all disciplines in the railways work together:

- The track layout must allow the local train to divert to a side track at a station, where it can slow down from approximately 80 km/h to standstill,
- Short block sections, cab signalling and advisory speeds must be used so that the overtaking intercity train can follow at the highest possible speed and at the shortest possible distance. This also requires precisely rescheduled paths; the intercity service will then pass the local train approximately 30 seconds after it comes to a halt, which is a normal duration for boarding and alighting of passengers.

- At this moment, the local train can close its doors; once this has been done, the intercity train will have passed and the local train is ready to depart.
- Immediate departure of the local train requires short block sections and cab signalling, so that it is not required to wait until the intercity train has cleared a whole block of approximately 1500 m. Since the local train still has a low speed, it requires only a short braking distance and one or two free short block sections would suffice.

The advantage is clear: with only 800 m additional track per station per direction and improved control, the throughput of a double track railway line with heterogeneous traffic can approach that of what is currently believed to be possible only on lines with four tracks. As can easily be seen, such approaches require that timetable designers, infrastructure designers, traffic control engineers and signalling engineers align their situation awareness, or their representation of how the system works and what it is capable of. It is also seen that a successful design can only be introduced once the overtaking process is decomposed into its subprocesses. Poor alignment between several functions implicitly will be paid for in the form of additional tracks that must be built, or unnecessarily long running times.

10.6.4 Measure how well processes work together

It is clear that a well-aligned co-operation between functions is essential. To see how well the functions in the cascade model work together and where improvements are possible, data must be gathered on the interface between functions. This thesis does not claim to provide a complete list, but a starting point could look like that in Table 10.2. This table shows the cascade model and how it can be measured whether each function hampers any other function.

Above the diagonal, interface performance indicators are given that allow the following functions to work properly. For example, if the (working) timetable design is properly done, it leaves few conflicts that must be solved during operations. Similarly, if the route setting process is performed well, this leads to many paths that are actually delivered conform the customer's demand, to few misroutings and to few unplanned signal stops. Below the diagonal, the table shows how frequently one of the lower functions fails to remain within the operational envelope provided by the higher level function. For example, a SPAD occurs when the driving process fails to remain within the limits of the route setting process; a delay occurs if any of the operational processes get outside the tolerance agreed on with train operating companies during timetable design. On the diagonal itself we find performance measures for each function itself, for example the number of routes that were set into a wrong direction or the number of times that a driver errs.

When interpreting these indicators they should not be allowed to lead a life of their own (e.g. avoiding doing rescheduling so that there will be no rescheduling errors, because that would result in a well-functioning performance indicator). The function of these indicators is to see how well the processes work smoothly together.

The measures can and should be related to each other. For example, the percentage of train runs that is involved in a SPAD should be related to the percentage of unplanned red signal approaches, thus allowing us to make statements about the driver's effectiveness in various situations. The amount of unplanned yellow signal approaches should be related to how well trains run on their paths and how well these paths were rescheduled in a conflict-free way. Frequent applications of emergency brakes by the automatic train protection system may indicate overreliance of drivers or an unreliable system with many false interventions, whilst

Table 10.2: Performance indicators on function interfaces.

	Infra design	Timetabling	Rescheduling	Route setting	Driving on path
Infra design	% superfluous infrastructure elements in design	restrictions for desired timetable design	restrictions for desired rescheduling	restrictions for desired route setting	Clearness of signalling information
Timetabling	trains scheduled that cannot be handled by infrastructure (too long, too heavy, wrong traction or train protection system)	% paths scheduled within customer acceptance % conflicts expected in operations	% conflicts to be solved in operations	% wrongly programmed routes	% incorrectly provided planned paths
Rescheduling	trains scheduled that cannot be handled by infrastructure (too long, too heavy, wrong traction or train protection system)	% delayed by unsolved conflicts	% paths rescheduled within customer acceptance % conflicts remaining in operations	% interlocking interventions	% incorrectly provided actual paths
Route setting	% routes set not in accordance with design intention (e.g. sending a train to a shunting signal)	% delayed by late route setting	% route unavailability	% wrong routes % no timely route	% paths delivered conform customer demand % misroutings % unplanned yellow aspects % unplanned red aspects
Driving on path	(only relevant via route setting)	% delayed by driving problem	% not running on path	% exceeding max. speed % SPAD	% running within agreed tolerance % emg. brake driver % emg. brake ATB % unnecessary brakings

frequent applications of emergency brakes by the driver himself may indicate that the signalling layout is not clear and thus provide clues about spots where a high SPAD probability is likely.

Collecting such data is technologically easily possible, either by analysing track occupation data automatically (Goverde et al. 2008) or by regularly reading the on-board monitoring equipment such as the Juridical Recording Unit (JRU, in Dutch: Automatische Ritregistratie). In aviation, such analyses are obligatory and, after each flight, data from the so-called Quick Access Recorder are stored in a database for further analysis (International Civil Aviation Organization 2005). In Dutch railways, the JRU is currently only examined when incidents happen, but in Italian railways (de Fabris et al. 2008) routine analysis is already practised. NS, as the biggest employer of train drivers in the Netherlands, does not systematically assess driver performance because of its 'big brother'-feeling. However, it would be very valuable to understand driver behaviour better for the population as a whole: for instance how the driving

style changes when changes to the ATB system are made, such as abandoning the cancel function or introducing ATB-VV (sections 9.4.2 and 9.4.3), or how frequently drivers rely on the ATB warning bell before they come into action, and how they exactly respond to yellow and red signals. It is necessary to collect such data on a large scale, in order to be able to identify whether and how the change affects driver behaviour. This would then allow the finding of reasoning errors such as those that were now made by design and management staff when such changes to train protection systems were made (see section 9.4).

Similar analyses might be worthwhile for the traffic control process, for example to learn how effectively traffic controllers reschedule in the case of problems, how often they face problems with the automatic route setting system et cetera. Such data can be used to improve processes and designs of support systems for operators.

Specific improvements for the traffic control process and train driving process will now be elaborated on in sections 10.7 and 10.8.

10.7 Unravelling the problem (C2): Traffic control improvements

10.7.1 Technology for supporting the human operator

For more than a hundred years, a prevalent ideal in railway engineering has been to correct human error through technological backups or automation. Although this technology-minded viewpoint certainly has led to the introduction of interlockings and train protection systems and thus to a situation in which human errors are rendered by and large harmless before accidents happen, this approach also has a drawback. Current safety systems generally check routine tasks such as braking commands (from the driver) or route setting commands (from the traffic controller). These are basically checks on ‘whether the right button has been pushed’. Since many of these lower level tasks have been automated or at least been made easier by use of technology, the human operator has become more and more a supervisor of the process; this holds both for drivers and traffic controllers. Sadly enough, human operators cannot perform well a supervisory process in which they are not actively involved. It easily leads to distraction and loss of skills and thus increases the human error probability in the (fewer) cases in which human action is required. By trying to remove the human factor from the equation, the danger is introduced that the cooperation between the human operator and the technology receives little attention and can even be made worse (Bainbridge 1987).

Instead of pushing buttons, the human operator’s task is now more that of a director of the process, who has to take decisions on a much higher level of abstraction. There are two reasons to consider a new allocation of functions between men and machine. Firstly, because human operators perform poorly in pure supervision tasks, it is required to keep the operator in the loop (Kaber and Endsley 1997). Secondly, technology can be used to support the operator’s awareness, e.g. by combining data from several sources so that it fits the more abstract tasks of the operator. The old paradigm, in which technology is always to be preferred above human beings, therefore needs to be changed – not because it is wrong, but because it is incomplete, so that other useful solutions remain out of the picture. The interlockings and train protection systems that have been achieved through the proven approach should, of course, not be thrown away. However, technology should not only be used to correct the operator, but also to assist him, for example by supporting his situation awareness. In this field, modifications in existing or future designs will be necessary. Designing to support operator situation awareness is already common in aviation (Endsley 1999) and should also become a main criterion to assess new designs in railways.

This requires situation awareness to be made measurable in a railway context. Golightly et.al. (2009) provide some preliminary trials of measuring situation awareness in a rail traffic control environment at different levels of automation, and discussion on how to measure situation awareness. The concept of situation awareness could be applied, or could have been applied, in several cases mentioned in this thesis, such as:

- the signalling system (Chapter 7); or
- the VPT system in several configurations (e.g. with/without conflict detection, or with/without automatic route setting; see Chapter 8. Isaksson-Lutteman et al. (2009) provide an example of a traffic control system that has been built to support situation awareness); or
- RouteLint (RouteLint is a current proposal to show the driver additional information from the traffic control system VPT so that he can see what is happening on the track ahead, for a more detailed description, see Appendix A7).

Endsley recommends to present data not on a low level format (i.e. making data that was coincidentally available on a separate screen) but preferably on a comprehension and projection level (i.e suitable for easy diagnosis), so that it can easily be seen whether the operator is about to achieve his goal. Several ways of developing comprehensible interfaces to enhance driver situation awareness are described by Young and Grenier (2009). Making comprehensible interfaces is meant to reduce the need for operators to scrape up data from several sources and combine it, which is something that can well be automated. Combining data from several sources makes the traffic situation more predictable and understandable, which enables operators to concentrate on finding solutions instead of losing time to gain an overview of the disturbed situation first. Cues that indicate the presence of prototypical situations are of prime importance.

10.7.2 DDA for solving conflicting infrastructure claims

The function of traffic control has intentionally become one of solving conflicting infrastructure claims, and consequently assigning each train a path by setting a route such that the service intention can be realised. This is mostly an anticipatory task and therefore it is necessary for a traffic controller to have a reliable estimation about when these conflicts will occur, so as to be able to detect them, diagnose their nature and act to resolve them.

Detect

Firstly, it is essential to have oversight about which track sections are assigned to trains, when and where trains need their route to be extended to, and when these blocked sections are expected to be available again for other trains. For a quick response, prototypical situations should be easy to identify for the traffic controllers. This allows them to see easily whether goals are, or will be, met. From the results so far, these prototypical situations comprise the following issues:

- violations of service intention;
- occurrence of failures and disturbances (primary delays);
- production conflicts (causes of secondary delays).

Further feedback from operators could expand this list. The current VPT system only shows failures and measured delays, so the other two would be worthwhile to add.

Diagnose

Not only are good estimates about the future system behaviour required, also the time taken to find a solution must be as short as possible; the shorter this time is, the less far the estimates need to go into the future and thus uncertainty in data can be reduced. This requires a good situation awareness and short communication times with other actors. Predictive displays can

help to show where problems exist and what the consequences of possible solutions are. The behaviour of the automatic route setting system should be completely understandable, which requires that it sets routes only in strict accordance with the traffic controller's plan (as is currently not the case) and it does not set routes for trains which still have an unsolved conflict under the responsibility of the traffic controller.

Act
Communication should be aimed at making decisions in cooperation with other controllers, not at asking what the situation actually is. Clearly visible envelopes can already reduce the need for active cooperation with other actors, since a controller knows that, if he remains within his envelope, he will not mess up other controller's processes. This might allow the use of result controls rather than action controls and can exploit the traffic controller's creativity. When changes are required to be made outside the operational envelope, relevant actors should also be informed. These are not only adjacent controllers, but also users of the infrastructure, such as drivers or maintenance gangs. In those cases where conflicts could not be solved, the direct conflict can still be avoided by slowing the train earlier such that it will not arrive before it can have its route set.

Suggestions for a support system
For improvements in the traffic controller's information systems, the blocking time model (see section 2.3.5) can also be used in operations (Pachl 2000; Kauppi et al. 2006; Pachl 2006). Use of this model is already common in railway scheduling in, for example, Germany, Switzerland and Sweden (Grayley 2005; Kauppi 2006; Hansen and Pachl 2008). By using the model in operations, it is possible to calculate conflict-free, deadlock-free schedules in real time (d'Ariano 2008). Although this is such a powerful tool that it may lead to operators who no longer can understand why a certain solution is optimal, this can certainly be a powerful tool to find conflicts in the actual schedule, show them to operators and suggest several possible solutions and their effects. Recent developments in predictive displays also allow the rescheduling task to be done manually in a quick and effective way (Kauppi 2006; Laube et al. 2007). Adopting one of these existing approaches could save time and money and provide a well-thought out display. It is recommended to explore several prototypical combinations of human – computer – interaction to find out what works best in the specific context of (Dutch) railway operation.

Although they can be powerful, the examples given above do not yet include the possibility to show conflicts with regard to the service intention. This functionality should therefore be added. Also, before copying one of the existing approaches, it should be made sure that these are indeed fully suitable for the Dutch situation and an explicit check for problems encountered in these foreign examples should be performed.

10.8 Unravelling the problem (C3): Train control improvements

10.8.1 DDA for timing of trains

The task of the driver is to operate the train such that it arrives at the planned time at its scheduled commercial stops, without hindering the traffic controller's ability to set routes for other trains and without violating safety-critical limitations. A train control system (regardless of whether the control task is performed by a human operator, by automation or by a combination of these) should therefore steer the train such that the driving process:
- complies with the *current plan* of the traffic controller (outcome of rescheduling process);

- stays within the *safety limits* as indicated by the signalling system (outcome of route setting process);
- can be *monitored* in its progress by the traffic controller (to allow conflict detection).

To transmit the traffic management information from the traffic control process to the train control process not only the message variety is important, but also the time it takes to transmit the message (i.e. the variety per unit of time) and the probability of a wrong message being delivered. For traffic management, also information about the train's status (location, speed, departure process, etc.) is required in order to re-schedule if necessary. While lineside signalling systems cannot provide such information to the traffic control process, newer signalling systems such as ETCS can (see Appendix A6).

Detect
In order for the driver to realise the path as planned by the traffic controller, he needs to be informed about changes in the reference path and he must be able to see whether he adheres to it. Since buffer times between two trains can be as small as quarter minutes, and this will be increasingly the case when utilisation is increased, drivers should ideally be able to see whether (and where) they are within a 15 second time window. This requires a much higher variety in information and a higher update frequency than can be provided by systems that are currently in use. However, today's technology could provide systems that are capable of doing so. It hardly needs be mentioned that this can only be done by providing the information in the cab.

Diagnose
The driver must be able to define a driving strategy, for example he must know when to start or finish certain subprocesses in the driving process. These are listed below, and again, those processes for which no information is currently being provided are underlined:

- when he will be able to depart;
- whether it is necessary to achieve the shortest possible running time;
- at what speed he should run in order to remain on his path or whether he is running on his path;
- whether he can decide to coast, and if so, at which point;
- where he should start to brake, given the braking characteristics of his train;
- where he should stop the train;
- in case of boarding and alighting: whether the whole train is standing along the platform.

We see that the current information systems hardly provide any of this information.

Act
Rail transport allows physically only for a limited number of actions that the driver can perform. Since the only useful actions are accelerating and decelerating (and stopping), the trick for effective and efficient driving is to know when, where and how strongly to apply these measures.

With regard to reaching the goal and keeping the train on its path, a new generation of traffic management systems should provide direct information on the timing of trains. This will enhance the realisation of the timetable and it can also keep the process away from the safety-critical boundaries.

10.8.2 DDA for safety-critical boundaries

Since (re)scheduling errors and timing errors in the driving process will always remain a possibility, as well as the fact that due to technical failures and environmental influences the

current situation may deviate from the planned situation, a backup system is required to prevent accidents.

Detect
Since the driver has to pay attention to several information sources, it is beneficial if he can decide himself at which moment he should look at which source, depending on task load and priority of tasks. Lineside signalling dictates that the driver should pay attention when a lineside message passes by (system-paced) and if he misses the information, there is no chance to recover from that. Therefore, not only traffic management information, but also safety-critical information should be available in the cab. In that case, the system is user-paced. This allows the driver to look again at the information when he is not sure what the message was; this is important because information in the human's short term memory will vanish within a few seconds unless it is paid attention to repeatedly (Encyclopædia Britannica Online 23 Jul. 2009b). Cab signalling also solves issues such as looking at the wrong signal or not being able to see the signal due to bad weather conditions or poor sighting distances. One might wonder why railways still use lineside signals, since cab signals were already commercially available before World War II. The quote below, coming from the supplier who also provided ATB system 30 years later, clearly illustrates this: '*Cab signals are the ultimate signals! (…) The increasing speed and density of railway traffic have created a demand for a more intensive method of signaling which the cab signal adequately provides. The recent trend toward high-speed streamlined passenger trains capable of developing a top speed of 110 miles an hour, is an indication of passenger train speeds in the near future (...). Cab signalling is the logical system to maintain the high-speed schedules of today and anticipate the requirements of the future*' (General Railway Signaling 1937). Since there are current plans to adapt the ATB system to allow speeds of 160 km/h (Götz 2008), this quote appears to be still relevant.

Diagnose
The signalling system only indicates a boundary that shows whether the current system state is safe or not safe; it therefore does not need as high a variety in its messages as traffic management information. Nevertheless, the signalling system should have a sufficiently high variety in its messages to clearly distinguish between all different situations which a driver can meet. We have seen several examples in which a 'smart' trick was invented so that the number of signal aspects was reduced to a number that is lower than the number of situations that the driver can meet. Examples are the Dutch yellow aspect, which gradually 'absorbed' also the meanings of other aspects such as 'short braking distance' or 'entry to dead end track' (section 7.3.3), or the UK example of the approach-controlled signal clearing (section 7.4.2). Ashby's Law says that the number of aspects must be as large as the number of messages that must be conveyed to the driver; as soon as several messages are combined in one aspect, this will sooner or later lead to confusion and wrong expectancies. The next section (10.8.3) will provide examples of how the driving process can be decomposed into a number of prototypical situations, which all require a distinct (cab or lineside) signal aspect so that the driver can easily see what he is expected to do. This therefore sets the requirement for the signal variety.

Act
With regard to safety-critical stops, the driver cannot produce much variety. The basic strategy is to stop the train, depending on the situation, either as soon as possible or at an accessible location, and ask for help (or wait until the restriction that caused the unplanned stop is over). If the train is stopped by the automatic train protection system, this comes as an

unplanned event so that the traffic controller should be informed, as rescheduling may be necessary.

10.8.3 Prototypical situations

The driver performs a safety-critical task; thus, the signalling system must be able to inform the driver unequivocally about the situation ahead. Several sources indicate that it would be useful to characterise typical situations in the driving process in order to help the driver to quickly diagnose the situation correctly. From the situation awareness literature, the importance of prototypical cues is made clear (Endsley 1999), whilst such approaches have (coincidentally or deliberately) been adopted in the design of ETCS operating modes (see Appendix A6) and in the previous Dutch signalling system (see section 9.4.4). To find prototypical situations, a good starting point is to make a functional decomposition of the driving process. Table 10.3 and Table 10.4 provide examples of such prototypical situations, called regimes. These lists are neither complete nor guaranteed correct, but they serve as

Table 10.3: Examples of driving regimes.

Regime:	Meaning:
No restrictions	Except for the speed limits originating in infrastructure and rolling stock capabilities, no restrictions apply. The driver should follow the planned path.
Reduced speed	Within speed limits due to infrastructure and rolling stock capabilities and route setting, the driver should follow the planned path.
Prepare to stop	The driver must prepare to stop. Following the planned path is of no importance. Speed should be reduced to x km/h; then, change to 'stop on sight' regime.
Stop on sight	The driver should regulate the speed such that he can stop within the distance over which he can oversee his route. Maximum speed x km/h.
Staff responsible	Moving the train is only allowed with permission of the traffic controller. This regime lasts until any other regime is signalled. (e.g. after unplanned stop, so that the route could have been revoked, or before a train is moved towards a signal showing 'stop').
Stop	Stop the train at the location which is indicated by the signal.
Emergency stop (danger in the train)	Stop the train as quickly as possible at a position that is accessible for emergency services. Call emergency services. (e.g. in the case of fire, need for medical help or criminal acts).
Emergency stop (danger for the train)	Stop as quickly as possible. Alarm traffic controller and other trains in the vicinity (e.g. in the case of loss of train integrity, loss of route integrity, derailments, SPADs, open doors or exceeding of loading gauge).

Table 10.4: Examples of halting regimes.

Regime:	Meaning:
Waiting	Keep the train stopped until one of the driving regimes applies.
Turn	Keep the train stopped. The train will continue into the opposite direction.
Departure	Notify passengers of departure. Close doors. Change to driving regime 'no restrictions' or 'reduced speed' or 'expect to stop'.
Station stop	When the train is next to a platform over its whole length, unlock doors. Keep the train stopped.
Parking	The train will be left unattended. Make sure that it cannot move during the time that it is unattended.

examples of what such a list could look like; this thesis aims to give a direction in which problems can very likely be solved, not the complete engineered solution itself.

Note that any of the driving regimes can only be allowed when none of the regimes lower in the list apply. So, if in doubt, an emergency stop is always applicable. This is well in line with current practices in railways, but can then be made explicit.

Further work is needed to make these lists complete. Other regimes can possibly also be found, either by following the functional decomposition or by reverse engineering several signaling systems. Each of the regimes could be indicated with a unique symbol; this in fact can and implicitly does form the basis of signalling systems. Given the functional approach, it should be possible in principle to find a corresponding signal aspect for each of the modes in every signalling system. It is beyond the scope of this thesis to do so, but it would be interesting to see whether this is possible and whether, for instance, the new European cab signalling system ETCS does support each of the regimes. Because the list serves mainly as an example, some 'difficult' regimes have been left out on purpose: for instance, (un)coupling, banking, shunting or 'reversing mode' (already existing in ETCS) could also be added.

For those readers who have knowledge of the Dutch signalling system, it can be easily seen that the yellow aspect indicates 'prepare to stop' if the previous signal indicated a driving regime that is higher in the list. In all other cases, yellow indicates 'stop on sight'. Thus, it can now easily be seen that the same aspect requires different behaviour in different situations.

So far, this section has provided several examples of how to improve. Although valuable, even more valuable is if the railway sector is a learning organisation so that many more of such improvements can be developed from within the companies. Thus, it is important that higher-level actors change their roles: it is necessary to make sure that operations function as intended and to ask for problems during operations to be reported – not to blame the one who caused or reported the problem, but to prevent the problem from re-occurring. The same approach can be used to solve safety problems and quality problems (Ellenbroek 2005). Much effort is probably required to change the culture throughout the entire rail industry to establish the trust that is necessary for these learning and feedback processes

10.9 Discussion

This section provides some discussion on the applicability of the results from this thesis. Section 10.9.1 discusses the usefulness of the theoretical framework; sections 10.9.2 - 10.9.5 discuss several issues with regard to practical implications that may arise when the railways are about to follow up the recommendations made earlier in this chapter.

10.9.1 Deriving improvements from the theoretical framework

The theoretical framework, consisting of the cascade model, the safe envelope and DDA sequences, connects functions in railway operation, information needs for each of the functions and the processes to exchange and obtain this information. The strong emphasis on functions and information flows make the framework generically applicable to a number of fields. This chapter has shown that this theoretical framework can be used for suggesting improvements: this was done to invert the emergent categories that were found using the grounded theory approach. It was hypothesised earlier in this thesis that the recommendations with regard to organisational learning and co-operation in the cascade model must be

implemented before structural improvements to traffic control and driver support systems can be made. This indeed is the case, since without a clear goal and designed process the organisation will not be able to direct the support systems such that they support the intended interaction. Nevertheless, this can be an iterative process since technological changes may provide possibilities that were not thought of at an earlier point in time. Once new technology has been introduced, feedback is required because otherwise the organisation is unable to learn whether the tools they provided have the intended effect.

10.9.2 Do technological developments indeed support the processes?

Although technological developments should support the production processes and should be derived from the intended processes, this is not always the case. In Table 10.5, an overview is given between the functions and technical solutions, which have been the subject of different studies in this thesis or its appendices.

We see that several technical changes do not lead to better supported processes (left column of the table), whilst at the same time many processes need better support (right column). We also see developments such as VPT (Chapter 8) and RouteLint (see Appendix A7), which were intended as an integrated control system but were reduced in their scope during their life. These projects resulted in a partial match between functions and technology. RouteLint (version 1.0), has its origins in the question how newer ICT solutions can help drivers and traffic controllers. In Table 10.5 this is called a coincidental hit, since this approach does not make sure that the driver is supplied with precisely the information he needs, but just with information that could be easily made available. This is not necessarily bad, but it will not produce advantages to their fullest extent. Given the success and enthusiasm about this first version, a second version is now being defined in which the information to be displayed plays a more central role. An intentional match can be found in the Dynamic Traffic Management solution for Schiphol (Appendix A5), where the process was designed first and the technology developed in order to enable this. Also the project of customer oriented logistics (Appendix A1) is on its way to becoming such an intentional hit, but much progress still needs to be made in order to get these principles work. The BB21 project had similar intentions, but shifted from goals to means, as was described in section 9.3.2. Also ETCS is mainly conceived from a technological viewpoint, since it can best be regarded as a standardised

Table 10.5: Classification of technological developments with examples.

		Process support realised?	
		No	**Yes**
Technological support intended	**No**	n/a	*Unsupported processes:* Conflict detection Path adherence
	Yes, technologically driven	*Technology-driven change (which are in fact also a shift from goals to means):* *ATB cancel function* Yellow symbol indicator displays P-signals Telerail → GSM-R ETCS	*Coincidental match:* RouteLint 1.0
	Yes, derived from process needs	*Shift from goals to means:* BB21	*Intentional match:* VPT (except conflict detection module which has not been realised) RouteLint 2.0 Customer oriented logistics Dynamic Traffic Management

European signalling and train protection system. Of course, this is quite a useful achievement, but it does not directly support the train control process in terms of realising the timetable. First user experiences in interviews indicate that the system's design is at the moment mainly useful for running a train over a main line, but the system is quite time-consuming for instance when shunting or reversing the train, for example because the on board computer must be re-started and data regarding the train and/or the driver (ID) have to be re-entered, which takes much time.

To prevent that any project or a proposal for a design change shifts from its goals to its means, it should be made sure that the team of designers regularly asks itself what the function is that they want to support and how exactly the user of their design is going to benefit from their work.

10.9.3 Add-on and front-end solutions

To solve SPAD problems, railways tend to look for add-on solutions that aim to bring a red signal approach to a safe conclusion. Possible measures can be aimed either at improving the driver's situation awareness or at the working of the automatic train protection. Examples are explicit driver education (so that they know about the gaps in the signalling system and how to cope with them), modifications to the signalling system, replacement of the signalling system by a cab signalling system such as ETCS, adding functionality to the train protection system such as ATB-VV in order to make trains stop automatically, or replacement of ATB by ETCS (network-wide or only at those spots where it is most necessary). All these measures can be justified by the argument that deviations from the planning can never be taken away completely, and troublesome situations will always be possible.

However, in addition to these add-ons, front-end solutions are also possible. The way to prevent accidents advocated in this thesis is not (only) through more safety systems, but most of all by preventing the traffic process from running into difficulties. This requires a more anticipative, process-oriented approach. Such an approach can also be used to improve performance, as anticipatory process control can solve problems taking more dimensions into account than the back-end solutions mentioned above. This will lead to fewer problems, although it can never be made sure that all problems will be solved. A nagging background problem is that if such front-end solutions manage to solve the vast majority of unplanned signal stops, drivers may get used to this and thus not focus their attention on signals, as they expect them to be green anyway. There is a trade-off between the likelihood of noticing a yellow signal aspect (which becomes lower when they are very rare) and a correct response to them once they have been noticed (which becomes higher when they are very rare). The current situation is typically one in which yellow signals are seen by drivers, but not given enough priority because they are seen too often, so that the intended meaning 'danger ahead' is now weaker than it should be. If this dilemma is to be solved, the question should be answered how well the control information is capable of keeping the process away from the safety-critical boundaries, and how big the probability is that distraction and/or wrong information make the driver unaware of safety-critical information. With answers to these questions, a net safety effect can be calculated.

A problem with add-on solutions is that they do not increase overall performance, because they act in too late a stadium: a train that has been stopped by the $(n+1)^{th}$ safety net is to be considered as in a safe state, but it does not contribute to quick and reliable transportation. Thus, these measures are costly without improving service quality. Several studies indicate that it is not worthwhile to invest huge amounts of money in rail safety as it does not prevent many casualties (ETSC 2003; Evans and Addison 2005). Although ETCS (or a similar

system) appears to be the best solution from a technical point of view, such a system is economically difficult to realise, at least if the benefits have to be based on safety improvement only. This leads to a very slow deployment of those systems, thus not offering the intended protection within a reasonable time.

A quicker deployment may be possible if these systems can not only contribute to improved safety, but also to other goals, such as improved quality of service (Onderzoeksraad voor de Veiligheid 2005a; Sierts et al. 2007). It is also possible to use ETCS elements for traffic management information, as is shown in Appendix A6. Application of ETCS is also recommended by consultancy firm Booz Allen Hamilton (2007) to improve capacity at pinch points in the Dutch rail network. ETCS can, basically in all its levels, increase the performance of the infrastructure, for example through use of short block sections and train-specific braking characteristics. A prerequisite for the latter is that the current, very conservative braking model be replaced by a more sophisticated one (Eichenberger 2007, Winter 2009). This will be solved with ETCS baseline 3.0.0. The benefits of (expensive) safety systems should therefore also be looked for and found in other areas than safety alone. This was already realised in 1920 when the originally separate Dutch railway companies had just merged and their different signalling systems had to be integrated. One of NS's board members indicated: '*If signal engineers would pay attention to the safety of traffic only, then they would perform their task wholly inadequately, because what matters is to implement safety measures with due regard for the demands of traffic, so that an accurate and speedy operation of traffic is not only possible, but also encouraged, yielding the most profitable result from the railway and its costly structure*' (Van Dijk 26-6-1920).

10.9.4 Dilemma of integration of control information in safety systems

Safety systems such as ETCS need time-consuming and expensive testing procedures, for good reasons. A possible disadvantage of integrating traffic management functions into them is that changes in the traffic management functions may also become delayed or expensive because of this exhaustive approval process. Many railways are now developing their own traffic management system to support drivers. Although understandable, this brings a number of risks.

- Firstly, the interoperability problems which are now almost solved for different traction energy systems and different train protection systems, will be replaced by a new problem of compatibility requiring that the driver's cab be fitted with different traffic management systems.
- Secondly, care should be taken when driver assistance systems can distract the driver's attention from safety-critical information for which no automatic train protection exists, in the Dutch case for example in the approach to red signals or at temporary speed restrictions (e.g. RouteLint, see Appendix A7). For the latter, no solution has yet been found. However, separate traffic management systems can surely be an option, and national developments like RouteLint could eventually be integrated into the European Timetable Displays as intended in UIC standard 612 (UIC 2008) or into the ETCS DMI at a later point.
- Finally, once the technology is available to display much more information to the driver, it is also tempting to display other (safety related) information such as maximum line speeds or temporary speed restrictions on these (cheaper) traffic management systems. However, if this information is required to meet the safety standards (SIL-4), and these systems have not been subject to the same tests as safety systems, such non-vital systems must not be used for such purposes.

Sometimes it is argued that safety systems are what ensure safety and that non-safety-related systems (such as a traffic management system) will therefore not influence the level of safety. However, this is not true, since this only considers the dilemma from a technical viewpoint and it presupposes error-free driver behaviour. As a consequence, also non-vital systems should be subject to a risk analysis if they are operated by employees who have safety critical actions in their responsibility. To help the driver in prioritising information, it may be necessary that safety systems can override traffic management systems.

From a theoretical viewpoint it is useful for the driver to have an integrated display with both the traffic management information and the safety-critical boundary (i.e. Movement Authority) so that he can have an overview of the situation in one view. However, pilot projects in railways such as RouteLint or the Swiss Co-Production project (for both, see Appendix A) indicate that it can also be useful to have these systems strictly separated, so that a driver clearly knows which information is safety-related and needs definitely be followed up (i.e. signalling information), and which information is additional (i.e. traffic management information). The main contribution of this thesis is that it is necessary to have both types of information; the precise way of showing these should be addressed in further research. However, in the case of a train protection system with full supervision (e.g. ETCS), the driver cannot exceed the operational envelope and an integrated picture can be shown.

10.9.5 The need to bring integrated solutions to a good conclusion

Many of the ideas that are currently being developed in the rail sector appear to be new. However, most of them are not, as they were already discovered in the past, but not brought to a satisfactory conclusion:

- Current plans of ProRail to increase capacity by another, more dynamic approach to timetabling and traffic control (Zoeteman et al. 2009), resemble those of the 'Transport Factory', a development of the old research and development department of NS (Maartens 1994).
- Co-production of traffic controllers and drivers by means of time windows as a clear communication tool is not new. Similar ideas were also discussed in the Netherlands in the 1990s (Poort et al. 1994; Weits 1997; Schaafsma 2001).
- The inadequacy of the yellow signal aspect was recognised in the mid 1980s (Middelraad 2000).
- Around 2003, ProRail was busy trying to develop itself towards becoming an HPO (which stood for High Performance Organization, but was meant to be the same as HRO) (Project Benutten & Bouwen 2003b; ProRail 2004). This development appears to have stopped for whatever reason. This research indicates that this development should be picked up again, but it also shows that it will be a long, hard road to travel to reach it, involving the rethinking of many fundamental concepts and assumptions in today's railway operation.

It has been shown that these approaches were promising from the viewpoint of the theoretical framework developed in this thesis. This raises the question as to how and why the developments mentioned above were actually not brought further. These could have ameliorated many problems that were found during this research. Knowing the answer to this question seems to be essential to avoid useful improvements being started but not being brought to a good conclusion.

One possible answer is that the unbundling of the integrated railways in 1995 led to fragmentation of knowledge and lack of common goals; whereas the developments mentioned above primarily have an integrated view. However, even if this explanation is true, it does not ask for a reunification of the railways, but it does seem to be necessary to overcome problems

that arise when one player has to invest so that other player(s) have bigger advantages. Such unbundling can also only be part of the answer: more recently, ProRail or one of its predecessors has often appeared to start the right developments, but then not always to finish them as they were originally intended. Examples are the efforts to become a High Performance Organization, developing Conflict Detection and Decision Support for traffic controllers and advisory speeds for drivers (Projectteam Definitiestudie VPT/BB 2002; Hemelrijk 2004), and the Het Spoor Meester project in its original, wide scope. It is apparent that ProRail is sufficiently smart to conceive the right solutions, but these are too often not implemented. Similar results were also found as a bottleneck in an earlier study (Schotanus and Zigterman 2005). Some practical implications of why innovations in railways may fail and what could be done to make them more successful are provided by Schaafsma (2004). He indicates the following reasons for failures:

1. Innovations take longer to be implemented than the time available given market constraints.
2. Managers tend to assign budget to the well-known current processes rather than to complex and uncertain innovation projects.
3. Having an effective marketing strategy for the innovation can be even more important than the innovation itself.
4. Even a company-internal innovation may require that the company behaves differently in the market.
5. Innovation requires learning from errors during the innovation process, but this is not a good approach to innovations for companies that cannot tolerate errors during daily operations; successful companies therefore try to improve gradually.
6. the most successful innovation may start as apparently pointless ideas, which causes managers not to support them initially.

Also, he cites six prerequisites to make innovations more successful:

1. The problem must be clearly formulated.
2. It must be clear who is primarily responsible for solving the problem.
3. The new vision must be disseminated, but without overloading stakeholders with information.
4. There must be an incentive for stakeholders to solve the problem.
5. Persistence and long-term commitment helps to overcome start-up problems.
6. Budget is required to implement the innovation; given the uncertainty in costs and benefits, it is required to make small steps so that the uncertainty can be reduced due to intermediate (test) results.

The answer to the question why useful improvements are finally not pursued is becoming an urgent one, since several useful projects are being pursued at this moment. These projects have evolved simultaneously more or less by coincidence, yet the framework developed in this thesis indicates that these developments should not be seen in isolation and should therefore be managed in a coherent way. These developments are:

- Key performance indicator ‘train path’, which leads to a clear goal to work on;
- Customer-oriented logistics, which leads to a coherent set of control loops for timetable design, traffic control and train control;
- New infrastructure design norms, based on functions to be fulfilled on the infrastructure.

A further enhancement to this set of projects could be a new project with the objective to develop smart infrastructure solutions for prototypical situations, to create better assistance systems for traffic controllers and drivers (which could fill in the principles of the customer-oriented logistics project), and a development to make the railways a learning organisation.

In summary, in order to make the introduction of improvements more successful, the railways should keep the following suggestions in their minds:
- A technological improvement will not automatically lead to a functional improvement for the user of that piece of technology, and can even work counterproductively if the precise operational processes are unknown;
- Improvements of safety systems should also be considered based on their potential to lead to improved capacity or quality of service;
- A driver has a safety-critical task, so that the introduction of driver assistance systems should always be accompanied by a safety analysis, also and especially if the system under consideration is likely to draw away the driver's attention from more critical tasks;
- There should be a concerted effort to find out why useful proposals for improvement are not brought to a satisfactory conclusion and which prerequisites must be fulfilled in order to make them successful.

10.10 Conclusion

This chapter has dealt with the question how process control can be improved to increase safety and quality in railways. The theoretical framework was helpful in doing so: from the grounded theory, twelve underlying causes for SPAD and quality problems were found, both at an organisational and an operational level. The theoretical framework, which consists of the cascade model, the safe envelope and DDA loops, permitted the coherent analysis of numerous developments, which helps to get a broader understanding of the problems under consideration, so that targeted improvements can be formulated. This chapter has provided the following main recommendations.

Firstly, the railway industry should adopt a framework such as the cascade model as a common reference and use this to apply the principles of a learning organisation. Problems during operations should be fed back so that an organisational learning process can be started or improved, to solve its problems. A clear goal that should be achieved as a prerequisite is to measure whether the process remains within the operational envelopes stemming from quality of service and safety requirements, in order to give a clear direction to the system's activities and motivate the operators working in it.

Secondly, the cooperation between the functions in the cascade should be better coordinated. Organisational learning is one of the ways to achieve this; but it is also necessary to create consistent flows of information between the levels in the cascade. From the safe envelope of operations, it can be seen that two types of function are required, each being a DDA sequence and each relating to one of the functions in the cascade model.
1. Control functions are meant for navigation within the operational envelope(s); and
2. Backup functions to make sure that the processes do not exceed the operational, and ultimately the viable envelope.

It was shown that not all of these functions are currently implemented. To bring rescheduling, train control and ultimately service quality in line, it is not only required that better support systems are developed or bought, but most of all that there is a change in paradigm. No longer should disturbances be denied and isolated, and the resulting problems primarily be 'solved' by backup systems designed to avoid unsafe consequences of erroneous control. Although such measures will remain necessary, disturbances should also be explicitly taken into account during timetable and infrastructure design, and in operations. These processes must cooperate to achieve their common goal, and these cooperative processes should be designed

before plunging into the (technological) solutions that support these processes. In short, the timing of trains should be controlled directly to use infrastructure as effectively as possible and to get the quality of service right.

Thirdly, decision-making for real-time operators should be enhanced, which requires not only shorter communication times, but also enhanced information systems that are capable of showing the operator's goal, to which extent this is reached, and the boundaries within which their processes should remain. In short, operators (both drivers and traffic controllers) must have accurate and compatible situation awareness. Situation awareness should be used as one of the assessment criteria for new operator support systems or procedures. Situation awareness can be improved by a clear identification of prototypical situations and clearly separated subtasks. It would be useful to integrate predictive and corrective design reviews in the design process, in order to be able to assess how new designs changes the operator's situation awareness.

11 Conclusions and recommendations

The main research question in this thesis was:
In what way could real time process control be improved to positively influence safety and quality of service in rail traffic operation?

To answer the research question, the following main steps were taken:
1. basic concepts from railway operation and safety science were analysed and gaps identified;
2. a framework, the so-called 'safe envelope of operations', was developed which can incorporate the main issues relevant to railway operation and safety science;
3. interviews and literature surveys were taken to identify to what extent current practices match the criteria from the framework;
4. options for improvement were explored.

The main results will be given below in more detail.

11.1 State of the art in research and need for additional knowledge

Both main players in the railway sector (ProRail and NS Reizigers) are still searching for their role and how to measure their performance. There is a tendency to assess performance more and more in terms of customer satisfaction and/or in indicators that relate to the actual service output to the customer. Since railways are a synchronised system, the timetable, which makes the synchronisation explicit, plays a central role for production of services. It was shown that 'better utilisation' has wider dimensions than just operating a higher number of trains on the same network: the system should avoid squeezing trains onto the network at the cost of actual customer demands. Increased utilisation also asks for making the system operate at a faster pace so that means of production are more quickly available for a next production cycle. Many functions cooperate to produce these services. Currently, these functions work in a strongly top-down-manner, with little bottom-up feedback. Additional information about feedback and co-operation between functions was indicated as a useful direction to create further understanding.

From safety science, it becomes clear that risks cannot be considered in isolation, but must be seen in the context of normal operations. However, most accident models consider safety separately from production goals. Also, existing models are not fully suitable to model the influence of process control on a dynamic system, such as railways. It was shown that a systemic model is required to be able to cope with 'production' integrated with 'safety', and to include technology, human factors and management.

11.2 Framework to understand process control, safety and quality of service

Cascade model

To provide an overview of the range of functions that are fulfilled to manage and execute railway operation, the cascade model (Figure 11.1) was introduced in this thesis. It shows the functions, the information flows between them, and the main input and output for the railway system. This ensemble of functions must produce railway services in safe way.

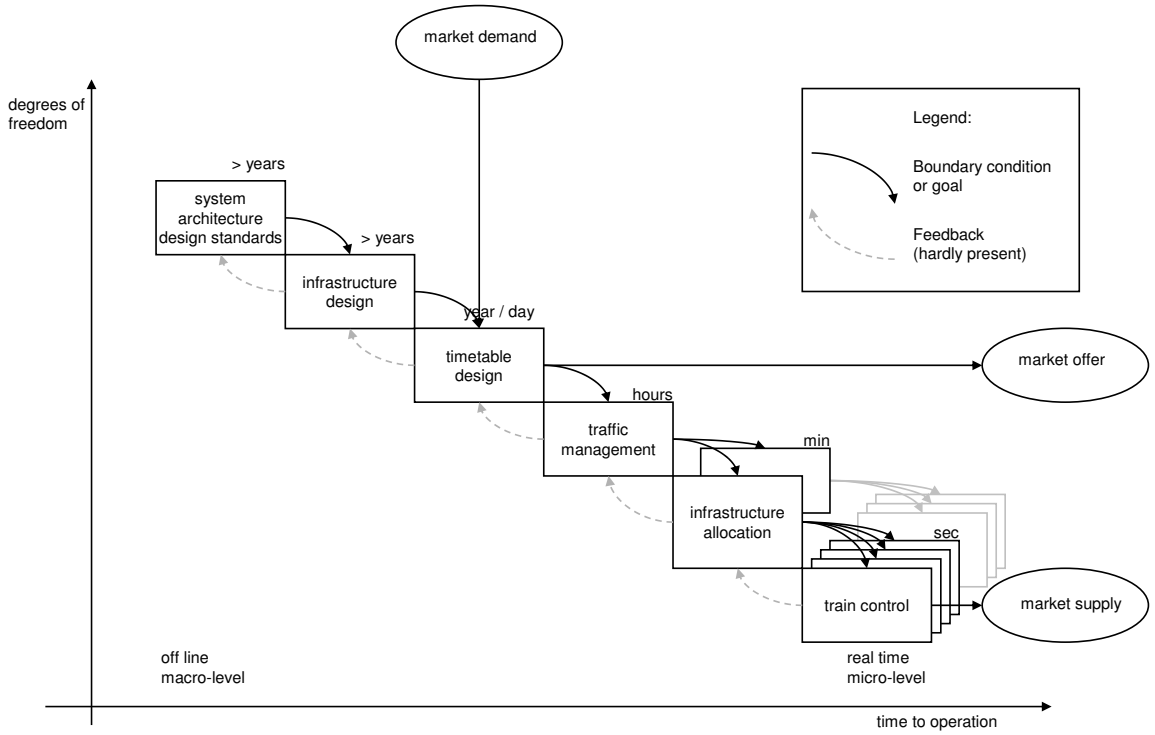


Figure 11.1: The cascade model.

Safe envelope of operations

To overcome the problems with current safety models this thesis developed its own theoretical framework. The safe envelope of operations can, on an abstract level, describe accident mechanisms in such a way that the drawbacks of existing safety models are alleviated. It makes clear within which boundaries the operational process must remain; not only the process itself, but also the boundaries can change over time which makes the model dynamic. It also provides a common language to describe the methods by which the operators in the system collaborate to stay within the operational and viable boundaries and to understand the failures in those control attempts. The framework indicates that actors who are controlling a system should be able to see information on the following items:

- What is the current system state?
- What is the goal and which tolerances are allowed (operational envelope)?
- Can the goal be reached in time (attainability envelope)?
- Where is the boundary of safe operation (viable envelope)?
- What is the ultimate moment to exercise control (controllability envelope)?

In distributed systems such as railways, the answer to these questions generally requires communication between different actors, which is therefore of vital importance to keep the system controllable and safe.

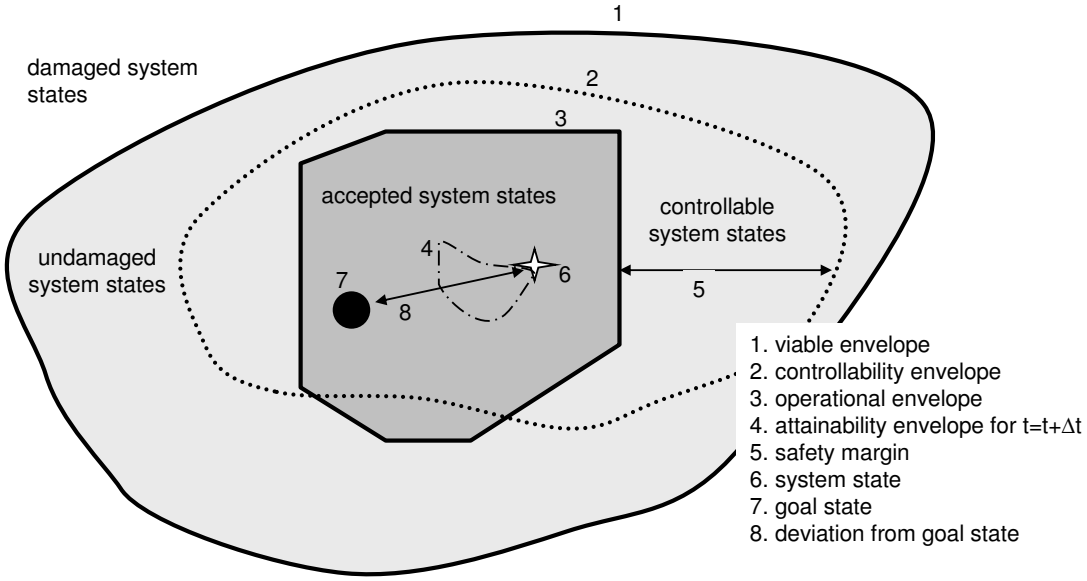


Figure 11.2: The model of the safe envelope, flattened to a two-dimensional image.

Detect – diagnose – act control sequence
Cybernetics, the general science for systems control and based on functions and information exchange between functions, made it possible to relate safety science, technology, human factors and management to each other. The basic concepts from these fields can all be represented as control loops which are represented as the detect-diagnose-act (DDA) sequences. This pushes forward the understanding of accidents in terms of control systems and it helped to formulate criteria for how to remain within the boundaries that the safe envelope model indicates. Also well known concepts such as buffer times in railways or barriers in safety science can be covered by this approach. Since effective control requires that control measures are taken in due time (which may be very short), it is preferable, if not essential, if regulators on any level have the ability to anticipate; also, it is preferable to keep the number of hierarchically placed regulators in a control system as low as possible, to avoid noise and delay due to additional communication.

The safe envelope model does meet the requirements, albeit for the moment only as a visualisation. This is an important first step that enables us to derive which variables are important to regulate. These stem from the objective of the system under consideration and the energy types in it and how these can be transferred to other elements inside or outside the system. Further steps, that require additional research, could be to try and model the envelope in a more formal mathematical way where possible; but this can be difficult since not all relations are easy to quantify. Another next step could be to think of ways to improve the visualisation of the multidimensional envelope, for instance by developing software that can

render cross-sections of various variables and show the dynamics of the way in which envelopes change.

11.3 Explanations for the current quality and safety problems in railways

Interviews with a range of actors from railway operations and design were carried out and combined with an analysis of professional and scientific literature to analyse current operations. This revealed much information with regard to the co-operation of functions and the extent to which actors have the information required for effective control (as mentioned above) at their disposal. This so-called grounded theory approach was a useful way to reduce a large amount of interesting, but blurry data to a manageable number of recurring phenomena that underlie the problems under consideration.

This grounded theory approach led to twelve related underlying causes for SPADs and performance problems, which are shown in Figure 11.3. These underlying causes can be grouped as three vicious circles:

A. Since no clear goal is set (1), it is not really clear for many operators what exactly they should achieve. This either results in self-formulated goals (3) and self-developed practices (10), or to operators clinging to their formal responsibilities (2). For several reasons, many operators are discouraged (7) and thus provide no feedback (8). Without feedback, there is insufficient organisational learning (9). Problems remain which further discourages operators.
B. Also, this causes operators to develop individual and therefore incompatible methods (10), which are partly a cause of inadequate control information and the loss of motivation. This starts a self-reinforcing process of operators retreating to 'islands' (12).
C. This also makes it difficult for designers to understand how they can support operators with better support systems, regulations etc (11). This again leads to a shift from goal to means (3) and support systems that do not always adequately provide the operators with the information they need (4,5,6)

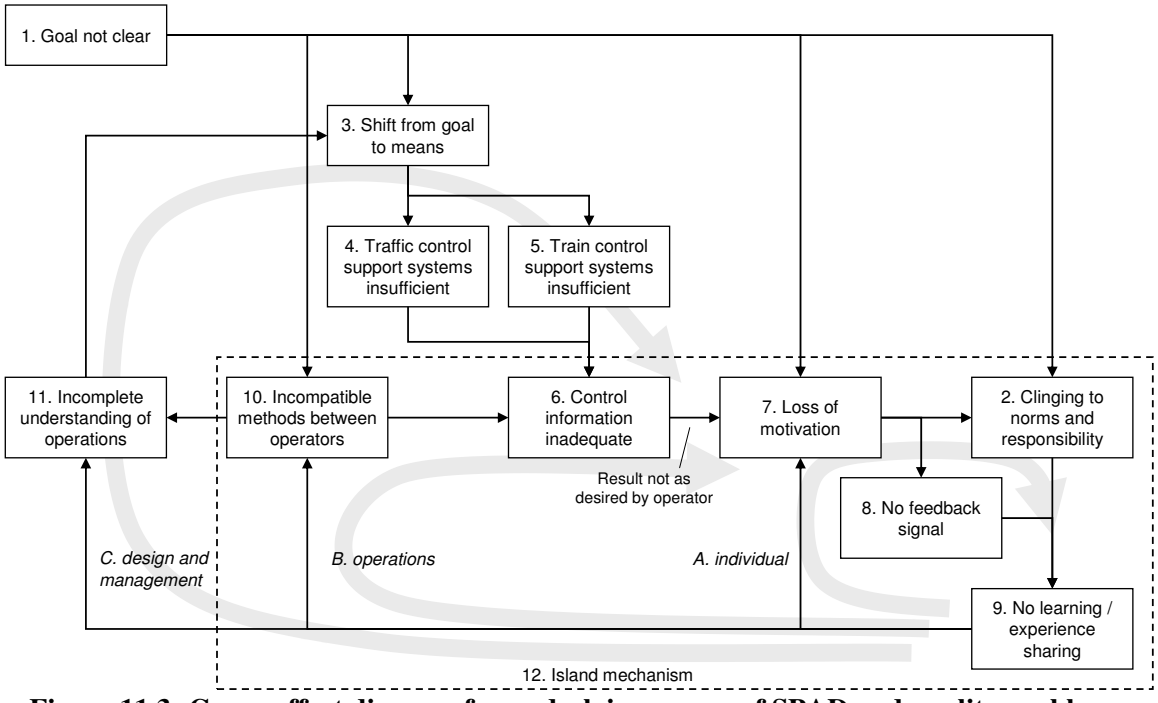


Figure 11.3: Cause-effect diagram for underlying causes of SPAD and quality problems.

The following underlying causes are especially worth mentioning:

- Since disturbances cannot be taken away, they must actively be counteracted. However, current infrastructure and timetable designs are made independently of the traffic control philosophy and without taking possibilities for effective traffic and train control tasks to cope with disturbances into account (10).
- It is a main task of the traffic controller to keep the system going even in the case of (internal or external) disturbances. Solving conflicts in the expected infrastructure use is mainly an anticipative task, but in the current situation, the process is difficult to predict (4, 6):
  - Information must be gathered from several sources and is in some cases incorrect;
  - Intended aids that would show the estimated future situation and expected conflicts have not been introduced;
  - The automatic route setting system behaves unpredictably in certain situations;
  - No updated, conflict-free, real time schedule is made in response to disruptions. Not having a feasible production schedule itself also makes the process unpredictable.

  As a result, the current process cannot be controlled on a detailed level more precisely than in whole minutes, and the railway system is slow to react to sudden disturbances. Currently, operator experience fills the gap of incomplete information, but this leads to different and in some cases incompatible working methods (10).
- Even though it is a main task of the driver to run according to schedule, this is currently not easily possible (5, 6):
  - Apart from the fact that new conflict-free schedules are normally not made in response to smaller disruptions, the driver is also not informed about possibly changed schedules;
  - He cannot see in sufficient detail whether he does adhere and how he can adhere better to the (re)scheduled path.

  This results in an unnecessarily high number of restrictive (yellow or red) signal approaches. Although timetable designers and traffic controllers rely on the driver to obey signals, imperfections in the route setting system, the signalling system and the train protection system, and the impossibility of faultless operator behaviour make this a dangerous assumption. There are several problems with the signalling system:
  - With lineside signals, the driver is provided with information only at discrete points, so that he must remember the signal aspect in some cases for a significant time.
  - Yellow, the most important signal aspect because it indicates the point where the driver should prepare to stop, is used to announce too many different situations. The driver is therefore not prepared optimally for a stop, because he only knows that he has to stop, but not precisely where.
  - A driver must be able to reliably relate a stop signal seen at some distance to his track, but it has been shown that this is not always possible.
  - The sighting distance of the stop signal itself is in some cases too small, so that the driver does not have enough time to respond and stop short of the signal.
  - Driver expectancy cannot be banned, since this is essential human behaviour. The focus should not be on prevention of expectancies, but effort should instead be spent to preventing that these expectancies can become wrong.
  - Route knowledge, which is in fact also an expectancy, is in some cases the last line of defence in order to assure safety. Given its weak formal meaning and the fact that the human memory works more reliably in response to external cues (and not to replace these), route knowledge should not be used as the only measure to ensure safe operations.

11.4 Improving risk control measures and their modelling in railways

With the theoretical framework and the underlying causes found with the grounded theory approach, it was possible to come up with new explanations for quality problems and SPAD incidents: both safety and quality can be seen as outcomes of a control process and occurring problems as failures of these. In order to improve, at least the following three groups of improvements should be taken into account, which will be elucidated below. Some changes should be made to the organisation as a whole; some to the cooperation of functions in the cascade model, and some to the support systems of traffic controllers and drivers.

Organisational changes
Firstly, the railways should become a learning organisation, so that they can and will find problems in their control processes themselves. Organisational learning implies that the organisation must actively look for unintended events during operations and encourage reporting of problems and errors. It is necessary to monitor whether renewed processes do lead to their intended effects, do not lead to harmful side-effects and whether they can be made more smoothly. This requires that experiences and behaviours from operation are fed back to designers and management, both before and after proposed changes. In Figure 11.4, this corresponds to performing better prediction and review at the design stage of the performance of the system elements and better analysis at the operational stage of measured data, leading to feedback of surprises and problems and suggestions for improvement to the appropriate levels. Also a clear goal should be set, so as to align the activities performed by the organisation, and the processes to achieve these goals should be designed before the (technological) solutions are chosen to support these processes.

Several useful projects that have adopted this approach have been started in the past. However, only a few of them have been finished in a way that the original objectives were realised. An essential question therefore is to find out why such useful improvements are

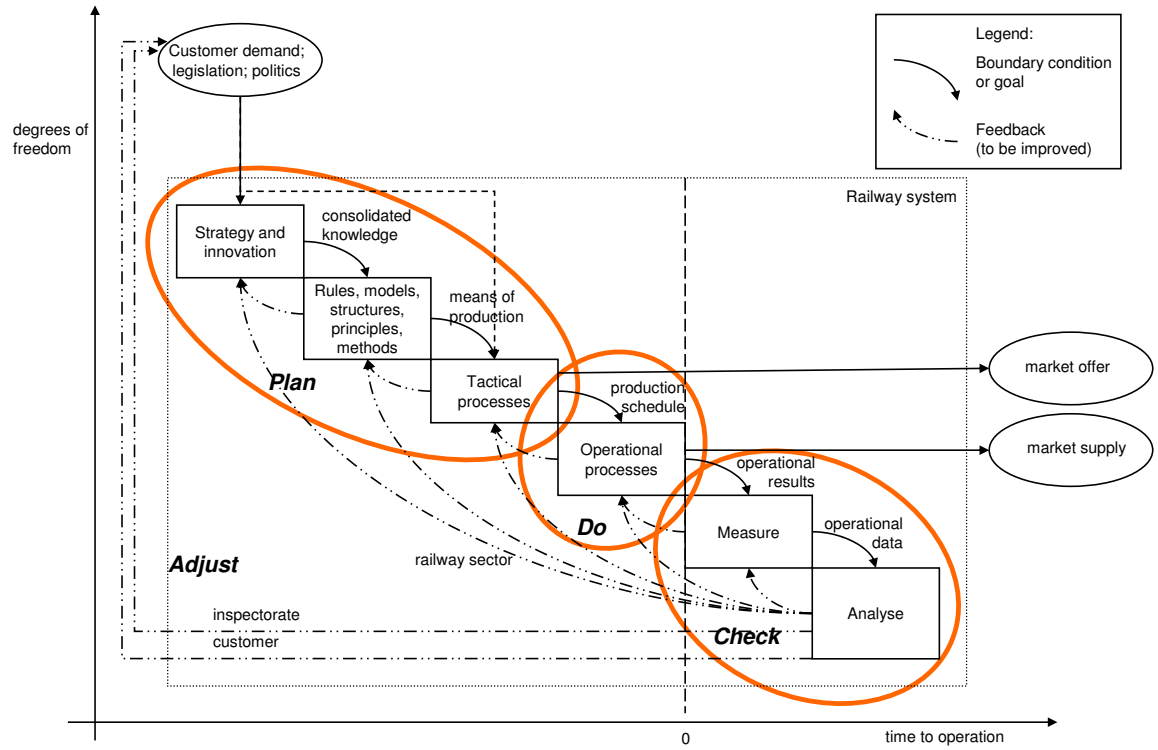


Figure 11.4: Extended cascade model.

finally not pursued to the end. This is an urgent question, since several useful projects are in progress at this moment. These projects have evolved simultaneously more or less by coincidence, yet they do cover a nearly complete package of useful and coherent developments and should therefore be managed in a coherent way. These developments are:
- Key performance indicator 'train path', which leads to a clear goal to work on;
- Customer-oriented logistics, which leads to a coherent set of control loops for timetable design, traffic control and train control;
- New infrastructure design norms, based on functions to be fulfilled on the infrastructure;

A further enhancement to this set of projects could be a new project with the objective to develop smart infrastructural solutions for prototypical situations.

Currently, structural performance analysis already takes place with regard to timetable design and multi-SPAD signals. It would be useful to extend such analyses also to other areas, such as signalling and infrastructure layout in general, and to operations support systems. The principles of a High Reliability Organisation provide a useful ideal situation to strive for.

Improved cooperation in the cascade model
Secondly, the functions in this cascade model should co-operate better, which requires coherent information flows and shared representations of how the system actually works. Organisational learning is a prerequisite to achieve this.

The railways currently try to combat process deviations by adding time buffers (within train paths as running time margins and between train paths as buffer times) and safety systems; however, both are costly and do not provide tools to steer the process such that deviations remain within acceptable tolerances. These can be specified by co-operating functions in the cascade model and represented as operational envelopes. It is recommended that the timing of trains is controlled directly, so that trains can be timed such that they will not normally arrive at yellow and red signals as long as these are not cleared. The control functions that should be supported for traffic management are:
A. Path definition
B. Conflict resolution and rescheduling
C. Route allocation
D. Remaining on path

The functions of conflict resolution, rescheduling and remaining on the assigned path are currently hardly supported.

This set of methods to bring rescheduling, train control and ultimately service quality in line, requires not only that better support systems are developed or bought, but most of all that there is a change in paradigm. No longer should disturbances be denied and isolated, but instead they should be explicitly taken into account during system design (including timetable design) and in operations. Disturbances should be prepared for and where necessary learned from.

Improve support system to enhance situational awareness
Thirdly, we have seen that drivers and traffic controllers lack important information to perform their tasks properly. Traffic controllers cannot easily see conflicting infrastructure claims and the consequences of possible solutions. Also, the automatic route setting system does behave unpredictably in some cases. This means that the situational awareness of traffic controllers is necessarily limited. Solving these issues will allow them to make better decisions.

Similarly, many SPAD incidents (about 36% of all SPADs in 2006) are attributed to drivers having an expectancy or being distracted, but the two concepts are linked and not clearly understood and defined – at least not in the context of Dutch railways. It was shown that having an expectancy is not a human weakness that leads to accidents, but a normal behaviour that enables human decision makers to guide their attention and to focus on those things that are expected to be important, which in turn leads to efficient behaviour. Both false expectations and distraction point towards insufficient situation awareness. The question should therefore not be how to get rid of driver's expectancies, but how to prevent them becoming wrong. Passing signals at danger is therefore not just a human factors or technical issue, but also an information exchange issue between traffic controllers reworking the plan and drivers executing it. New support systems are required to empower them and make it easier to find an adequate solution in a short time. The safe envelope can identify which information this is; the cascade model can be used to see which functions could provide such missing information. Technology should not only be used to replace human operators, but also to help them. Situation awareness should become one of the main criteria to assess whether a technological change is an improvement for operator behaviour.

11.5 Recommendations

By the use of a theoretical framework consisting of a mix of methods from safety science, engineering, human factors and social sciences, it was possible to find causes and solutions that go further than the explanations for SPADs (Signals Passed At Danger) and delays that are currently found by the railways. The following main changes are proposed:

1. The railways should adhere to the principles of a learning organisation, predicting the performance of its designs (both technological, procedural (timetable) and organisational and monitoring their resulting performance; this requires actively looking for problems in order to learn from them and being open for reporting of operational surprises by staff. Thus, the railway industry is made fit to find problems during operations and solve these adequately.
2. The co-operation of functions as described in the cascade model should be explicitly taken into account, to avoid optimisation of sub-processes such that this hampers other processes or to avoid that this even works counterproductive. In the tactical phase (e.g defining the required means of productions and how these will be used), the possibilities to cope with delays and disturbances during operations should be explicitly taken into account.
3. In cooperation between traffic control and drivers, the timing of trains should be controlled directly, so that trains remain within acceptable tolerances. Thus, trains will generally not arrive at signals before these could be cleared, and unplanned stops and SPAD risks will be reduced. This requires that support tools for drivers and traffic controllers are renewed accordingly. Situation awareness should be a main assessment criterion to see how a proposed change influences the operator's ability to understand the situation.

11.5.1 Recommendations for railway operations

Under the first heading below, several recommendations are summed up to make the railway organisation fit for improvements. Under the second and third heading, some more concrete grips are provided for improvements. A warning that should be given at this point, is that it will not suffice to simply make the technical changes described here without changing the organisation as well; similar problems will then re-emerge sooner or later.

Recommendations for organisational learning

- Use the theoretical framework, consisting of the cascade model, the safe envelope model and the DDA sequences to help understand mechanisms behind quality and safety issues from a dynamic perspective and by taking all relevant actors into account.
- Adopt a new philosophy in which disturbances during operations are not seen as unwelcome events, but as part of reality that should be catered for during operations. This requires changes to the design philosophy of infrastructure and timetables. The aim should be to provide a robust service output which can be achieved through more flexible operations, since the availability of more options at the operational level provides more opportunities to keep the process within the constraints as given by the external function of the timetable (e.g. the service intention);
- Use situation awareness as a criterion to assess whether technological changes are improvements to human operators.
- Include situation awareness-related entries in SPAD checklists, so that information can be retrieved in a standardised way with respect to situation awareness problems.
- Find out why many projects narrow their scope on the fly and thus do not deliver the originally intended outcome.
- Extend performance analyses such that not only includes timetabling issues are adjusted, but also for example the signalling and infrastructure layout, the effectiveness of traffic control (e.g. behaviour of traffic control staff and the automatic route setting system) and the behaviour of drivers. To do so, collect and analyse operations data, such as which signal aspects are displayed in signals, how many conflicts remain during operations, how often backup systems intervene, etc.
- Introduce an explicit design assessment and review program to predict both the desired and the unwanted results of changes, and to steer the design process to ensure it meets its objectives. To prevent shifting from goals to means, it should be made sure that the team of designers regularly asks itself what the function is that they want to support and how exactly the user of their design is going to benefit from their work.
- Make sure that there is an organisation in the railway industry that can check whether the rules are still correct.

Recommendations for improved control

- Prevent trains from running into yellow signal aspects by better traffic management and better adherence to the planned timelines of the train.
- Define prototypical situations to be encountered during the tactical and operational processes in the cascade diagram through functional breakdowns of processes, and develop a set of standard solutions for each of them.
- Find out which functions route knowledge is supposed to have in supporting these prototypical situations and consider whether route knowledge is also the best possible way to fulfil these functions. Formalise the definition, function, provision and assessment of route knowledge.
- Adapt the automatic route setting system such that it will not set routes to shunting signals for train movements.
- ETCS can be deployed considerably quicker if it is also used for other purposes than safety alone. ETCS on board equipment gives opportunities for several functions in the cascade model and should be explicitly steered in its final design and implementation to fulfilling these:
  - better infrastructure and timetable design (ability to make short block sections relatively cheaply, which makes shorter headways possible);

- traffic management (both to improve path adherence of trains and to provide detailed data to reschedule trains);
- for better signalling (with the required variety and offering cab signalling at all application levels, thus eliminating visibility and interpretation problems);
- better train protection (full supervision).

Recommendations for improved backup

- If no cab signalling will be introduced within a short span of time, make the existing signalling system clearer by increasing the message variety and by making the relation between the track and its corresponding stop signal more clear.
- Bring the approach speed of red signals, their sighting distances and the required braking distance in line.
- Solve the ambiguous meaning of the yellow signal aspect. Adding additional aspects to differentiate between different situations will help drivers to adjust their expectancies.
- Pay explicit attention to the signalling problems during the train driver training.

11.5.2 Recommendations for further scientific research

With regard to the safe envelope of operations, it is suggested that further research is carried out into the possibilities to quantifying and thus further specifying it.

- It may be useful to formulate the relationship between several variables in a mathematical way, but this can be difficult for other variables, for example when human factors are involved or when the effect of gradually crumbling knowledge must be quantified to reflect changes in one of the envelopes. Explore the extent to which expert judgement can reduce these difficulties.
- Another useful step could be to think of ways to improve the visualisation of the multidimensional envelope, for instance by developing software that can render cross-sections of various variables and show the dynamics of the way in which envelopes change.
- There are two essentially different paradigms with regard to how control information and safety-critical information should be displayed to the driver. From one viewpoint it is useful for the driver to have an integrated display, so that he can have an overview of the situation in one view. From another viewpoint, it can also be useful to make a clear division so that a driver always knows which information is safety-related and needs definitely be followed up (i.e. signalling information), and which information is additional (i.e. traffic management information). It would be useful to go deeper into this dilemma and see whether and how it could be solved.
- In generic terms, the theoretical requirements used in this thesis are not different for maintenance gangs or level crossing users than for train movements, so that the safe envelope could be applied to their processes as well; however, the functions and the restrictions under which these functions must be executed may be different so that other safety measures may result. Further analysis along the lines of this thesis is needed to establish this.

A last recommendation is to remember that operations cannot work without clear support from management, but that management cannot work properly without knowledge of actual operations.

References

Achilles, A. (2008) Die Migration von ETCS aus dem Blickwinkel des Produktlebenszyklus, in: *Signal + Draht* (10): pp. 666-670.

Air affairs (2006) *Route Knowledge Study*, RSSB, London.

Al-Ibrahim, A. (2008) *SMD als dynamische TAD - onderzoek van traject Utrecht-Gouda*, Universiteit van Amsterdam, Amsterdam.

Albrecht, T., J. van Luipen, I. A. Hansen and V. A. Weeda (2007) Bessere Echtzeitinformationen für Triebfahrzeugführer und Fahrdienstleiter, in: *Eisenbahningenieur* 58(6): pp. 73-79.

Apel, W. (1968) Linienförmige Zugbeeinflussung bei den Niederländischen Eisenbahnen, in: *Signal und Schiene* (7): pp. 294-296.

Argyris, C. and D. A. Schön (1996) *Organizational learning II - theory, method, and practice.*, Addison-Wesley, Amsterdam.

Ashby, W. R. (1956) *An introduction to cybernetics*, Chapman and Hall, London.

Aulin-Ahmavaara, A. Y. (1979) The law of requisite hierarchy, in: *Kybernetes* 8: pp. 259-256.

Avelino, F., M. C. G. te Brömmelstroet and G. Hulster (2006) *The Politics of Timetable Planning: Comparing the Dutch to the Swiss*, Colloquium Vervoersplanologisch Speurwerk 2006, Amsterdam, Stichting CVS.

Aven, T. (1992) *Reliability and Risk Analysis*, Elsevier Applied Science, London.

Baas, A. (2007) *Deelrijwegen en capaciteit - gevolgen van het mogelijk niet meer gebruiken van deelrijwegen*, ProRail, Utrecht.

Bainbridge, L. (1987) Ironies of automation, in: J. Rasmussen, K. Duncan and J. Leplat (eds.) *New technology and human error,* John Wiley & Sons, pp. 271-283.

Balfe, N., J. R. Wilson, S. Sharples and T. Clarke (2009) *Effects of level of signalling automation on workload and performance*, 3rd international rail human factors conference, Lille.

Barter, W. A. M. (2008) *ERTMS Level 2: effect on capacity compared with "best practive" conventional signalling*, Computers in Railways XI, Toledo, WIT Press.

Baud, F. M. (1984) Procescomputers en ontregelingen: mogelijkheden en beperkingen, in: *Openbaar vervoer* 17(10): pp. 261-265.

Bayes, T. (1763) An essay toward solving a problem in the doctrine of chances, in: *Philosophical transactions of the Royal Society of London* (53): pp. 370-418.

Beekman, J. (2008) *Evaluatie van het leer- en verbetersysteem van ProRail - Waar staat ProRail als organisatie nu?*, ProRail, Utrecht.

Beer, S. (1985) *Diagnosing the system for organisations*, John Wiley & Sons, Chichester.

Besluit Spoorwegpersoneel (2004), 's Gravenhage.

Bischof, W. (1986) Ein neues Signalsystem bei den SBB, in: *Schweizer Eisenbahn-Revue* (1): pp. 28-34.

Boeije, H. (2005) *Analyseren in kwalitatief onderzoek - denken en doen*, Boom Onderwijs, Amsterdam.

Booz Allen Hamilton (2007) *Study into best practice measures capacity enhancement for railways*, Booz Allen Hamilton, London.

Bouwman, M. and D. Middelkoop (2001) *SIMONE toetst dienstregelingen*.

Bovy, P. H. L. and C. D. Van Goeverden (1994) *De rol van kwaliteit in het personenvervoersysteem gemeten aan de verplaatsingstijdfactor*, Openbaar Vervoer Colloquium, Utrecht.

Branton, P. (1979) Investigations into the skills of train-driving, in: *Ergonomics* 22(2): pp. 155-164.

Brehm, J. and S. Gates (1997) *Working, shirking, sabotage: bureaucratic response to a democratic public*, University of Michigan Press, Ann Arbor.

Busby, J. S. and R. E. Hibberd (2002) Mutual misconceptions between designers and operators of hazardous systems, in: *Research in Engineering Design* (13): pp. 132-138.

Carstens, P. F., F. Bruijn, J. van den Top and H. de Raadt (2007) *Onderzoek incident zijdelingse aanrijding na STS sein 76 (d.d. 16 juni 2007) te Weesp, in relatie tot 8 voorgaande STS passages*, ProRail, Utrecht.

Casti, J. L. (1985) Canonical Models and the Law of Requisite Variety, in: *Journal of optimiaztion theory and applications* 46(4): pp. 455-459.

Christensen, C. M. (2006) The ongoing process of building a theory of disruption, in: *The Journal of Product Innovation Management* (23): pp. 39-55.

Conant, R. and W. R. Ashby (1970) Every good regulator of a system must be a model of that system, in: *International Journal of Systems Science* 1(2): pp. 89-97.

Corning, P. A. (2001) Control information - the missing element in Norbert Wiener's cybernetic paradigm?, in: *Kybernetes* 39(9-10): pp. 1272-1288.

d'Ariano, A. (2008) *Improving real-time train dispatching: models, algorithms and applications*, TRAIL, Delft.

Dabekaussen, M. (2005) *On track for reducing SPAD's - An analysis of the root causes of SPAD's using PRISMA Rail*, TU Eindhoven, Eindhoven.

Davis, D. R. (1966) Railway Signals Passed at Danger: the Drivers, Circumstances and Psychological Processes, in: *Ergonomics* 9(3): pp. 211 - 222.

De Bruijn, D. (2007) *Zichtproef MatrixSignaalGever*, Intergo, Utrect.

de Bruijne, M. (2006) *Networked reliability - institutional fragmentation and the reliability of service provision in crtitical infrastructures*, TU Delft, Delft.

de Bruijne, M., M. J. G. van Eeten, E. Roe and P. R. Schulman (2006) Assuring high reliability of service provision in critical infrastructures, in: *International journal of critical infrastructures* 2(2/3): pp. 231-246.

de Fabris, S., G. Longo and G. Medeossi (2008) *Automated analysis of train event recorder data to improve micro-simulation models*, Computers in Railways XI, Toledo, WIT Press.

de Groot, A. D. (1961) *Methodologie - grondslagen van onderzoek en denken in de gedragdswetenschappen*, Mouton&Co, 's Gravenhage.

de Leeuw, A. C. J. (2001) *Bedrijfskundig management - primair proces, strategie en organisatie*, Van Gorcum, Assen.

De Raadt, H. (2008) *Rood seinpassages - rijden voorbij rood sein verminderen door aanpassing van de infrastructuur*, SPHOV / ProRail, Utrecht.

de Vos tot Nederveen Cappel, H. A. E., R. van Dijk, F. J. Jansen, W. Jansen, G. Verheul and J. J. de Heer (1964) Seinwezen en telecommunicatie, in: *Spoor- en tramwegen* 37(jubileumnummer): pp. 88-96.

de Wit, J. and H. van Gent (1996) *Economie en transport*, Uitgeverij Lemma, Utrecht.

DeBlois, L. (1926) *Industrial safety organization for executives and engineer*, McGraw-Hill Book Company, New York.

Deming, W. E. (1982) *Out of the crisis*, MIT, Cambridge, Massachusetts.

Department for Transport (2007) *Rail Technical Strategy*, DfT, London.

Deutsches Institut für Normierung e.V. (DIN) (1994) *DIN V 19250 - Leittechnik; grundlegende Sicherheits-betrachtungen für MSR-Schutzeinrichtungen*, Beuth Verlag, Berlin.

Dijksma and Van Hijum (2006) *Motie van de leden Dijksma en Van Hijum*, Tweede Kamer der Staten-Generaal, 's Gravenhage.

Dijkstra, G. (2004) *Veiligheid: "een veilige balans met kosten"*, ProRail, Utrecht.

Dijkstra, G. (2007) Voor spoorinnovaties staat het sein al jaren op rood, in: *OV-Magazine* (may).

Directive 2001/14/EC *allocation of railway infrastructure capacity and the levying of charges for the use of railway infrastructure and safety certification)*, European Parliament, Strassbourg.

Dunn, N. and A. Williamson (2009) *Mitigating the effects of monotony: the role of task complexity*, 3rd international rail human factors conference, Lille.

Eichenberger, P. (2007) Kapazitätssteigerung duch ETCS, in: *Signal + Draht* (3): pp. 6-14.

El Koursi, E. M., S. Fletcher, L. Tordai and J. Rodriguez (2006) *SAMNET Synthesis Report - Safety Management and Interoperability*, SAMRAIL Consortium, Villeneuve d'Ascq.

Ellenbroek, A. B. T. (2005) *Veiligheid 'in control' - Een referentiekader voor de control van veiligheid*, Universiteit van Amsterdam, Amsterdam graduate Business School, Executive Master of Internal Auditing, Amsterdam.

Elvik, R. (2005) *Laws of accident causation*, 18th ICTCT workshop, Helsinki, International Cooperation on theories and concepts in traffic safety.

EN13816 (2002) *Transportation - Logistics and services - public transport*, European Committee for Standardization, Brussels.

Encyclopædia Britannica Online (01 Sep. 2009) "computer science", [http://www.britannica.com/EBchecked/topic/130675/computer-science]

Encyclopædia Britannica Online (23 Jul. 2009a) "attention", [http://www.britannica.com/EBchecked/topic/42134/attention]

Encyclopædia Britannica Online (23 Jul. 2009b) "memory", [http://www.britannica.com/EBchecked/topic/374487/memory]

Endsley, M. R. (1995) Toward a theory of situation awareness in dynamic systems, in: *Human Factors* 37(1): pp. 32-64.

Endsley, M. R. (1999) *Situation Awareness and Human Error: Designing to Support Human Performance*, High Consequence Systems Surety Conference, Albuquerque, NM.

Endsley, M. R. and E. O. Kiris (1995) The out-of-the-loop performance problem and level of control in automation, in: *Human Factors* 37(2): pp. 381-394.

ETSC (2003) *Cost effective EU transport safety measures*, Brussels.

Eurlings, C. M. P. S. (6 april 2009) *Informatie over een aantal zaken rond de prestaties en activiteiten van NS en ProRail*, 's Gravenhage, pp. KST129799.

European Commission (2006) *Commission decision concerning the technical specification of interoperability relating to the subsystem 'Traffic Operation and Management' of the trans-European conventional rail system*, 2006/920/EC, Brussels.

European Railway Agency (2009) *ETCS Driver Machine Interface*, ERA, Valenciennes.

Evans, A. W. and J. D. Addison (2005) *Interactions between rail and road safety*, Imperial College, London.

Farrington-Darby, T. and J. R. Wilson (2006) The nature of expertise: A review, in: *Applied Ergonomics* (37): pp. 17-32.

Farrington-Darby, T. and J. R. Wilson (2009) Understanding social interactions in complex work: a video ethnography, in: *Cognition, Technology & Work* (11): pp. 1-15.

Farrington-Darby, T., J. R. Wilson and B. J. Norris (2005) Investigating train driver behaviour: the use of lineside information when regulating speed, in: J. R. Wilson, B. Norris, T. Clarke and A. Mills (eds.) *Rail human factors: supporting the integrated railway,* Ashgate, Aldershot, pp. 60-69.

Farrington-Darby, T., J. R. Wilson, B. J. Norris and T. Clarke (2006) A naturalistic study of railway controllers, in: *Ergonomics* 49(12): pp. 1370-1394.

Fay, A. (1999) *Wissensbasierte Entscheidungsunterstützung für die Disposition im Schienenverkehr*, VDI-Verlag, Düsseldorf.

Fischer, R. (2004) *Het spoor naar een duurzame samenleving - Een onderzoek naar de knelpunten in de spoorsector die een transitie naar een duurzame mobiliteit belemmeren*, Technische Universiteit Eindhoven, Eindhoven.

Franklin, G. F., J. D. Powell and A. Emami-Naeini (2006) *Feedback control of dynamic systems*, Prentice-Hall, Upper Saddle River.

Garland, D. J., J. A. Wise and V. D. Hopkin (1999) *Handbook of aviation human factors*, Lawrence Erlbaum, Mahwah, NJ.

Gelders, D., M. Galetzka, J. P. Verckens and E. Seydel (2008) Showing results? An analysis of the perceptions of internal and external stakeholders of the public performance communication by the Belgian and Dutch Railways, in: *Government Information Quarterly* (25): pp. 221-238.

General Railway Signaling (1937) *G-R-S Coded Cab Signals*, GRS, Rochester, USA.

Gibson, J. J. and L. E. Crooks (1938) A theoretical analysis of automobile-driving, in: *American Journal of Psychology* 51(3): pp. 453-471.

Gigerenzer, G. and D. G. Goldstein (1996) Reasoning the Fast and Frugal Way: Models of Bounded Rationality, in: *Psychological Review* 103(4): pp. 650-669.

Glaser, B. G. and A. Strauss (1967) *The Discovery of Grounded Theory - Strategies for Qualitative Research*, Aldine Publishing Co, Chicago.

Golightly, D., N. Balfe, S. Sharples and E. Lowe (2009) *Measuring situation awareness in rail signalling*, 3rd international rail human factors conference, Lille.

Götz, W. W. J. (2008) *ATB Code Groen - haalbaarheidsonderzoek*, Inspectie Verkeer en Waterstaat, Utrecht.

Gouweloos, R., R. Jonker and P. Regtuijt (2003) *Verkenning Besturingsconcept Railverkeersleiding*, ProRail, Utrecht.

Goverde, R. M. P. (2003) *PRRING II - Inzicht in de betrouwbaarheid van realisatiegegevens treinactiviteiten*, ProRail, Utrecht.

Goverde, R. M. P. (2005) *Punctuality of Railway Operations and Timetable Stability Analysis*, Delft University of Technology, Delft.

Goverde, R. M. P., W. Daamen and I. A. Hansen (2008) *Automatic identification of route conflict occurences and their consequences*, Computers in Railways XI, Toledo, WIT Press.

Grant, R. (2003) Da Vinci will manage Spain's first 350 km/h line, in: *Railway Gazette International* (August 2003): pp. 495-496.

Grayley, C. (2005) Smoothing Sweden's PATH planning, in: *Railway Gazette International* (October 2005): pp. 639-640.

Griffioen, E. and J. R. Vorderegger (2001) *RvTV studie stoptonende seinen, versie 0.9; RnV/01/T42.004.100*, Railned Spoorwegveiligheid, Utrecht.

Groeneweg, J. (2002) *Controlling the controllable - preventing business upsets*, Global Safety Group, Leiden.

Grunwald, A. J., J. B. Robertson and J. J. Harfield (1981) Experimental evaluation of a perspective tunnel display for three-dimensional helicopter approaches, in: *Journal of Guidance and Control* 4(5): pp. 623-631.

Haddon, W. (1973) Energy damage and the ten countermeasures, in: *Human Factors* 15(4): pp. 355-366.

Hahn, J. and T. F. Edgar (2002) Process Control Systems, in: R. A. Meyers (ed.) *Encyclopedia of Physical Science and Technology,* Academic Press, San Diego, pp. 111-126.

Haig, B. D. (1995) Grounded Theory as Scientific Method, [http://www.ed.uiuc.edu/eps/pes-yearbook/95_docs/haig.html]

Hale, A. R. (2005) *Veiligheidsmanagement: wat weten we zeker?*, NVVK Veiligheidscongres, Eindhoven, NVVK.

Hale, A. R., B. J. M. Ale, L. H. J. Goossens, L. J. Bellamy, M. L. Mud, I. A. Papazoglou, A. Bloemhoff, H. Baksteen, J. Post and J. I. H. Oh (2007) Modeling accidents for prioritizing prevention, in: *Reliability Engineering & System Safety* 92(12): pp. 1701-1715.

Hale, A. R. and I. Glendon (1987) *Individual Behaviour in the Control of Danger*, Elsevier, Amsterdam.

Hale, A. R. and M. Hale (1970) Accidents in perspective, in: *Occupational psychology* (44): pp. 115–122.

Hansen, I. A. (2001) *Improving railway punctuality by automatic piloting*, IEEE Intelligent Transpot Systems Conference Proceedings, Oakland (CA) - USA, pp. 792-797.

Hansen, I. A. (2003) Waar blijft de razendsnelle stoptrein - betere benutting met metroachtig treinmaterieel, in: *Verkeerskunde* (8).

Hansen, I. A. (2004) *Increase of capacity through optimised timetabling*, Computers in Railways IX, WIT Press, pp. 529-538.

Hansen, I. A. (2006) *State-of-the-art of Railway Operations Research*, Computers in Railways X, WIT Press, pp. 565-579.

Hansen, I. A. and J. Pachl (2008) *Railway timetable & traffic*, Eurailpress, Hamburg.

Happel, O. (1959) Sperrzeitentreppe als Grundlage für die Fahrplankonstruktion, in: *Eisenbahntechnische Rundschau* 8(2): pp. 79-90.

Heinrich, H. W. (1941) *Industrial accident prevention - a scientific approach*, McGraw-Hill Book Company, New York.

Hemelrijk, R. (2004) *De Groene Golf (internal presentation)*, ProRail, Utrecht.

Henley, E. J. and H. Kumamoto (1981) *Reliability Engineering and risk assessment*, Prentice-Hall, London.

Heylighen, F. and C. Joslyn (2001) Cybernetics and second-order cybernetics, in: R. A. Meyers (ed.) *Encyclopedia of Physical Science & Technology (3rd ed.),* Academic Press, New York.

Hill, C., H. Blachard and M. Carey (2007) Informing signage positioning rules through a human factors assessment of signal/sign co-location, in: J. R. Wilson, B. Norris, T. Clarke and A. Mills (eds.) *People and rail systems: human factors at the heart of the railway,* Ashgate, Aldershot, pp. 203-214.

Hinzen, A. (1993) *Der Einfluss des menschlichen Fehlers auf die Sicherheit der Eisenbahn*, RWTH Aachen, Aachen.

Hofstra, K. (2005) *Eindrapportage systeemkosten en -opbrengsten van het Nederlandse spoorwegnet*, ProRail, Utrecht.

Holleman, J. (1991) Van irriterende naar intermitterende dodeman in SGM, in: *Matblad* (3): pp. 29-30.

Hollnagel, E. (2001) *Barrier Analysis And Accident Prevention*, Roissy CDG, France.

Hollnagel, E. (2004) *Barrier Analysis And Accident Prevention*, Ashgate, Aldershot.

Hollnagel, E., C. P. Nemeth and S. Dekker (2008) *Resilience Engineering Perspectives: remaining sensitive to the possibility of failure*, Ashgate, Aldershot.

Hooghiemstra, J. S., D. M. Middelkoop and M. J. G. Teunisse (1998) *Improving the Dutch railway services by network-wide timetable simulation*, Computers in Railways VI, Lisbon, WIT Press.

Houtenbos, M. (2008) *Expecting the unexpected - A study of interactive driving behaviour at intersections*, TRAIL, Delft.

Huisman, T., R. J. Boucherie and N. M. van Dijk (2002) A solvable queueing network model for railway networks and its validation and applications for the Netherlands, in: *European Journal of Operational Research* (142): pp. 30–51.

Human Engineering (2006) *Who, other than the driver, has a role in reducing SPADs?*, Rail Safety & Standards Board, Westbury-On-Trym.

Inspectie Verkeer en Waterstaat (2004a) *Op 20 maart 2003 vindt om 11:46 uur te Roermond een botsing plaats tussen een goederentrein en een reizigerstrein*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2004b) *Op 21 mei 2004 vindt om 18:32 uur te Amsterdam Centraal een botsing plaats tussen een leeg materieeltrein en een reizigerstrein*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2004c) *Op 27 maart 2004 omstreeks 13:00 uur rijdt trein 2249 ter hoogte van station Den Haag=Mariahoeve voorbij stoptonend sein 184*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2004d) *Seinveiligheidstoring te Rotterdam-Lombardijen op 20 augustus 2003*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2005a) *Op 30 september 2004 vindt om 17:46 uur een botsing plaats tussen een reizigerstrein en een locomotief te Roosendaal.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2005b) *Op donderdag 8 april 2004 om 17:50 uur rijdt trein 5558 voorbij stoptonend sein 1084 te Utrecht=Overvecht*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2006) *STS-passages 2005 - oorzaak, gevolgen en trends over de periode 2001 - 2005*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2007a) Nieuwe beveiliging op Amsterdam – Utrecht: onderhuidse revolutie, in: *Toezicht op Railveiligheid* juni.

Inspectie Verkeer en Waterstaat (2007b) *Op 12 maart 2007 om 18:26 uur rijdt te Amsterdam Muiderpoort een dubbeldekker zonder reizigers voorbij stoptonend sein en botst zijdelings tegen een intercity*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2007c) *Op dinsdag 21 november 2006 om 10.10 uur botst te Arnhem een goederentrein frontaal tegen een reizigerstrein.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2007d) *Op maandag 20 november 2006 om 10:00 uur botst te Rotterdam Centraal een rangeerdeel tegen een goederentrein.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2007e) *Op vrijdag 30 juni 2006 ontspoort om 15:32 uur een reizigerstrein ter hoogte van Zwammerdam (spoor Bodegraven - Alphen a/d Rijn)*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2007f) *STS-passages 2006 - analyse en resultaten over de periode 2002 - 2006*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2007g) *Toezicht in veilige handen*, IVW, Den Haag.

Inspectie Verkeer en Waterstaat (2008a) *Op 7 april 2008 ontspoort om 16:57 uur in de metrotunnel bij Centraal Station te Amsterdam een metrotrein*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008b) *Op 12 februari 2008 komen om 19:34 uur, na een stoptonend sein passage, twee treinen tegenover elkaar tot stilstand tussen Barneveld Centrum en Lunteren.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008c) *Op 26 november 2007 om 14:19 uur rijdt een reizigerstrein in de Schipholtunnel voorbij een gedoofd sein, dat rood behoort te tonen.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008d) *Op donderdag 26 juli 2007 komen om 11:05 uur, na een stoptonend sein passage, twee treinen vlak tegenover elkaar tot stilstand te Tilburg West.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008e) *Op donderdag 29 maart 2007 rijdt om 16:59 uur te Harmelen aansluiting een reizigerstrein met hoge snelheid voorbij een stoptonend sein. Een naderende goederentrein kan op tijd stoppen.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008f) *Op zaterdag 16 juni 2007 om 22:15 uur rijdt een reizigerstrein voorbij een stoptonend sein op het emplacement Weesp, waarna een zijdelingse botsing plaatsvindt met een andere reizigerstrein.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008g) *Op zondag 18 november 2007 rijdt om 00:10 uur een intercity te Sittard voorbij een stoptonend sein en komt bijna in botsing met een andere intercity*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2008h) *STS-passages 2007 - Analyse en resultaten over de periode 2003-2007*, IVW, Utrecht.

Inspectie Verkeer en Waterstaat (2009a) *Onderzoek opleiding en wegbekendheid van machinisten*, IVW, Utrecht.

Inspectie Verkeer en Waterstaat (2009b) *Op zaterdag 11 oktober 2008 vindt om 11:02 uur te Gouda een zijdelingse aanrijding plaats tussen een intercitytrein van NS Reizigers en een internationale trein van Thalys Nederland.*, Inspectie Verkeer en Waterstaat, Utrecht.

Inspectie Verkeer en Waterstaat (2009c) *STS-passages 2008 - Analyse en resultaten over de periode 2004-2008*, IVW, Utrecht.

International Civil Aviation Organization (2005) *ICAO Accident Prevention Programme*, ICAO, Montréal.

IRSE (1995) *European Railway Signalling*, A&C Black, London.

Isaksson-Lutteman, G., A. Kauppi, A. W. Andersson, B. Sandblad and M. Erlandsson (2009) *Operative tests of a new system for train traffic control*, 3rd international rail human factors conference, Lille.

ISO8402 (1986) *Quality vocabulary*, International Standards Organisation, Geneva.

ISO9000 (2000) *Quality management systems - Fundamentals and vocabulary*, International Organization for Standardization, Geneva.

Jacobs, J. (2008) Rescheduling, in: I. A. Hansen and J. Pachl (eds.) *Railway timetable & traffic,* Eurailpress, Hamburg.

Jagtman, H. M. (2004) *Road safety by design - a decision support tool for identifying ex ante evaluation issues of road safety measures*, Eburon, Delft.

Jansson, A., E. Olsson and L. Kecklund (2005) Acting or reacting? A cognitive work analysis approach to the train driver task, in: J. R. Wilson, B. Norris, T. Clarke and A. Mills (eds.) *Rail human factors: supporting the integrated railway,* Ashgate, Aldershot, pp. 40-49.

Johnson, W. G. (1973) *MORT - the management oversight and risk tree*, US Atomic Energy Commission, Washington.

Jongerius, R. T. (1993) *Spoorwegongevallen in Nederland 1839-1993*, Schuyt & co, Haarlem.

Kaber, D. B. and M. R. Endsley (1997) Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety, in: *Process safety Progress* 16(3): pp. 126-131.

Kaplan, S. and B. J. Garrick (1981) On the quantitative definition of risk, in: *Risk Analysis* 1(1): pp. 11-27.

Kauppi, A. (2006) *A Human-Computer Interaction Approach to Train Traffic Control*, Uppsala University, Uppsala.

Kauppi, A., J. Wikström, B. Sandblad and A. W. Andersson (2006) Future train traffic control: control by re-planning, in: *Cognition Technology and Work* 8(1): pp. 50-56.

Kecklund, L., M. Ingre, G. Kecklund, M. Söderström, T. Akerstedt, E. Lindberg, A. Jansson, E. Olsson, B. Sandblad and P. Almqvist (2001) *The TRAIN-project: railway safety and the train driver information environment and work situation - a summary of the main results*, 2. Signalling Safety, London.

Kingston, J., R. Nertney, R. Frei, P. Schallier and F. Koornneef (2004) *Barrier Analysis Analysed in MORT Perspective*, PSAM7/ESREL '04 International Conference on Probabilistic Safety Assessment and Management, Berlin, Germany.

Kirwan, B. (1994) *A guide to practical human reliability assessment*, Taylor&Francis, London.

Kjellén, U. (2000) *Prevention of accidents through experience feedback*, Taylor & Francis, London.

Klein, S. (2004) *Alles Zufall - die Kraft, die unser Leben bestimmt*, Rowohlt, Reinbek bei Hamburg.

Klein, S. (2006) *Zeit: Der Stoff, aus dem das Leben ist. Eine Gebrauchsanleitung*, Fischer S. Verlag, Frankfurt am Main.

Klerk, B. J. (9 July 2004) Meer veiligheid maakt de trein onbetaalbaar, in: *de Volkskrant*.

Klerk, B. J. (31 March 2003) *Strategie Treinbeveiliging en -beheersing*, ProRail, Utrecht.

Klerk, B. J. and R. J. In 't Veld (2005) Van kijken naar luisteren: modern toezicht op spoorveiligheid, in: Kraan et al. (ed.) *Van het toezicht, de veiligheid en de wetgeving,* SDU, Den Haag, pp. 79-87.

Kletz, T. A. (2006) Accident investigation: keep asking "why?" in: *Journal of Hazardous Materials* (130): pp. 69-75.

Knapp, K. (1998/1999) Driver expectancy, traffic control and roadway design, in: *Technology News* (December - January): pp. 1-2.

Kohls, H.-H. and R. Watson (2008) *SPAD - reducing time table risk*, Computers in Railways XI, Toledo, WIT Press.

Koopman, D. (2007) *Vervoersprestatieverbetering van spoorwegen tijdens calamiteiten*, TU Delft, Delft.

Laube, F. and M. Lüthi (2008) Making every second count, in: *Railway Gazette International* 164(12): pp. 952-954.

Laube, F., S. Roos, R. Wüst, M. Lüthi and U. Weidmann (2007) PULS 90 - Ein systemumfassender Ansatz zur Leistungssteigerung von Eisenbahnnetzen, in: *Eisenbahntechnische Rundschau* 2007(3): pp. 104-107.

Lenior, D., W. Janssen, M. Neerincx and K. Schreibers (2006) Human-factors engineering for smart transport: design support for car drivers and train traffic controllers., in: *Applied Ergonomics* 37(4): pp. 479-490.

Lenior, T. M. J. (1993) Analyses of cognitive procesees in train traffic control, in: *Ergonomics* 36(11): pp. 1361-1368.

Leveson, N. G. (2002) *System safety engineering: back to the future*, Massachusetts Institute of Technology, Cambridge, MA.

Leveson, N. G. (2004) A new accident model for engineering safer systems, in: *Safety Science* (42): pp. 237–270.

Lundberg, J., C. Rollenhagen and E. Hollnagel (2009) What-You-Look-For-Is-What-You-Find – The consequences of underlying accident models in eight accident investigation manuals, in: *Safety science* in press.

Luther, R., H. Livingstone, T. Gipson and E. Grimes (2007) Understanding driver route knowledge, in: J. R. Wilson, B. Norris, T. Clarke and A. Mills (eds.) *People and rail systems: human factors at the heart of the railway,* Ashgate, Aldershot, pp. 71-78.

Lüthi, M. (2008) *Evaluation of energy saving strategies in heavily used rail networks by implementing an integrated real-time rescheduling system*, Computers in Railways XI, Toledo, WIT Press.

Lüthi, M. (2009) *Improving the efficiency of heavily used railway networks through integrated real-time rescheduling*, ETH Zürich, Zürich.

Lüthi, M., U. Weidmann, F. Laube and G. Medeossi (2006) *Rescheduling and Train Control: A New Framework for Railroad Traffic Control in Heavily Used Networks*, 86th Transportation Research Board Annual Meeting 2007, Washington DC.

Maarschalkerweerd, M. S. P. A. M. (1997) *Knooppuntdienstregelingen - overzicht van Nederland, Duitsland en Zwitserland*, RailNed, Utrecht.

Maartens, M. (3 July 2009) Ai! Alweer de trein gemist, in: *NRC Next*: pp. 20-21.

Maartens, M. (1994) De Vervoerfabriek - hogere frequenties, compactere stations en meer flexibiliteit, in: *Openbaar Vervoer* (12): pp. 53.

Makkinga, F., H. Berends, M. Deerenberg and P. Regtuijt (2002) *Business scenario's voor CSBO*, Holland Railconsult, Utrecht.

Martens, M. H. and M. R. J. Fox (2007) Do familiarity and expectations change perceptions? Drivers' glances and response to changes, in: *Transportation Research Part F* (10): pp. 476-492.

Mazzarello, M. and E. Ottaviani (2005) *A traffic management system for real-time traffic optimisation in railways*, 1st International Seminar on Railway Operations Research, Delft.

Merchant, K. A. and W. A. van der Stede (2003) *Management Control Systems - Performance measurement, evaluation and incentives*, Prentice Hall, Harlow, England.

Middelraad, P. (2000) *Voorgeschiedenis, Ontstaan en Evolutie van het NS-Lichtseinstelsel*, NS Railinfrabeheer, Utrecht.

MIL-STD-882D (2000) *Standard Practice for system safety.*, US Department of Defense, Washington, DC.

Millen, L., S. C. Sharples, D. Golightly and N. Balfe (2009) *The application of cognitive work analysis in rail*, 3rd international rail human factors conference, Lille.

Ministerie van Verkeer en Waterstaat (29 november 2004) *Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden (KVR21734)*, SDU, 's Gravenhage.

Ministerie van Verkeer en Waterstaat (2004a) *Nota Mobiliteit; naar een betrouwbare en voorspelbare bereikbaarheid*, Den Haag.

Ministerie van Verkeer en Waterstaat (2004b) *Veiligheid op de rails - tweede kadernota voor de veiligheid van het railvervoer in Nederland*, Den Haag.

Ministerie van Verkeer en Waterstaat (2007) *Actieplan Groei op het spoor - Uitwerking korte termijn kabinetsambitie*, Ministerie van Verkeer en Waterstaat, 's Gravenhage.

Montigel, M. (2008) Innovatives Bahnleitsystem optimiert den Zugverkehr im Lötschberg-Basistunnel, in: *Signal + Draht* 100(9): pp. 18-22.

Moser, P. (2004) Planen im magischen Dreieck, in: C. Kräuchi and U. Stöckli (eds.) *Mehr Zug für die Schweiz - Die Bahn 2000-Story,* AS Verlag, Zürich.

Mulholland, A. and A. L. Macaulay (2006) *Architecture and the Integrated Architecture Framework*, CapGemini, London.

N.V. Nederlandsche Spoorwegen (1964) *Trein- en Rangeerdienstreglement (C5201)*, NS, Utrecht.

N.V. Nederlandsche Spoorwegen (1968) *Seinstelsels en beveiligingen bij N.S.*, Nederlandsche Spoorwegen, Utrecht.

N.V. Nederlandsche Spoorwegen (1973) *Spoorwegwetgeving - Deel II*, Nederlandsche Spoorwegen, Utrecht.

N.V. Nederlandsche Spoorwegen (±1952) *Cursus voor opleiding tot technisch ambtenaar - Het seinreglement*, NS, Utrecht.

Nederlandse Spoorwegen (2 March 2007) *Expertmeeting 'Menselijke fouten op het spoor - Recht doen aan mens, bedrijf en maatschappij'*, Nederlandse Spoorwegen, Utrecht.

Nederlandse Spoorwegen (1969) *Spoor naar '75: Plan voor de toekomst van NS in de Nederlandse Samenleving in de jaren 1970-1975*, Nederlandse Spoorwegen, Utrecht.

Nederlandse Spoorwegen (1978) *Meer over de relaisbeveiliging*, Nederlandse Spoorwegen, Utrecht.

Nederlandse Spoorwegen (1988) *Rail 21 - sporen naar een nieuwe eeuw*, NS, Utrecht.

Nederlandse Spoorwegen (1989) *C5504/II - Voorschriften voor de bediening van wissel- en seininrichtingen - Deel II Automatische Trein-Beïnvloeding*, Nederlandse Spoorwegen, Utrecht.

Nederlandse Spoorwegen (1993) Van zeventien eilandjes naar één duidelijk produkt verkeersleiding, in: *Kadans* (7/8): pp. 14-16.

Nederlandse Spoorwegen (2006) *Vervoerplan NS 2006*, Utrecht.

Nederlandse Spoorwegen (2008) *NS Jaarverslag 2007*, Utrecht.

Nederlandse Spoorwegen (2009a) *NS Jaarverslag 2008*, Utrecht.

Nederlandse Spoorwegen (2009b) *Vervoerplan NS 2009*, Nederlandse Spoorwegen, Utrecht.

Nederlandse Spoorwegen (november 2003) Beter benutten van spoor is enorm rendabel, in: *Trens* (7): pp. 23-25.

Nederlandse Spoorwegen, Railion and ProRail (2006) *ATB++ Systeemdefinitie SRS versie 1.93*, Utrecht.

Neerincx, M. A. and E. Griffioen (1996) Cognitive task analysis - harmonizing tasks to human capacities, in: *Ergonomics* 39(4): pp. 543-561.

NS Reizigers (2006) *Veiligheidsmanagementsysteem*, NS, Utrecht.

NS Reizigers (2008) *Veilig rijden met leeg materieel - Analyse en aanbevelingen voor het terugdringen van STS-passages*, NS, Utrecht.

NS Verkeersleiding (2000) *Evaluatie locatieproef CS/BO Utrecht IV*, Utrecht.

OECD (1997) *Road safety principles and models - review of descriptive predictive risk and accident consequence models*, OECD, Paris.

Oetting, A. (2008) Zuglaufregelung - optimierte Steuerung der Züge im Betrieb, in: *Eisenbahntechnische Rundschau* (10): pp. 654-658.

Onderwater, P. and R. v. Leusden (1999) Zonetrein vult gat in NS-dienstregeling, in: *OV Magazine* 5(7): pp. 14-17.

Onderzoeksraad voor de Veiligheid (2005a) *Door rood op Amsterdam CS*, Onderzoeksraad voor de Veiligheid, Den Haag.

Onderzoeksraad voor de Veiligheid (2005b) *Ontsporing goederentrein bij Apeldoorn op 30 april 2003*, Onderzoeksraad voor de Veiligheid,, Den Haag.

Pachl, J. (2005) Entwicklung der Leit- und Sicherungstechnik für das System Bahn, in: *Eisenbahntechnische Rundschau* 54(3): pp. 96-102.

Pachl, J. (2008) Die Bedeutung betrieblicher Regelwerke für die Leit- und Sicherungstechnik, in: *Signal + Draht* 100(12): pp. 32-38.

Perrow, C. (1984) *Normal accidents - living with high-risk technoligies*, Basic Books, New York.

Pessers, V. (2007) *What did you expect? - A case study about the critical succes factors for passenger transport companies in the Netherlands*, Tilburg University and Arcadis Infra, Tilburg / Amersfoort.

Peters, B. (2006) analyse drgl 2006 vs 2007, [http://finance.groups.yahoo.com/group/OVL/message/165603]

Pidgeon, N. F., B. A. Turner and D. I. Blockley (1991) The use of Grounded Theory for conceptual analysis in knowledge elicitation, in: *International Journal of man-machine studies* 35: pp. 151-173.

Plamann, V. and H.-F. Klindt (2008) Hoher Automatisierungsgrad bei der Hamburger Hochbahn, in: *Signal + Draht* 100(10): pp. 10-15.

Poort, J. (1992) ATB Nieuwe Generatie, in: *Het matblad* 1992(5).

Poort, J., J. Postmes, R. Vissers and M. Wink (1994) *Functioneel concept BB21*, Railned Innovatie, Utrecht.

Poort, J. P. (2002) *Grenzen aan benutting*, Nyfer, Breukelen.

Project Benutten & Bouwen (2003a) *Benutten en Bouwen - het plan van de spoorsector*, Utrecht.

Project Benutten & Bouwen (2003b) *Plan van aanpak Ontwerpopdracht Dienstregeling 2007*, Utrecht.

Projectteam Definitiestudie VPT/BB (2002) *Operationele Concept Definitie (OCD) VPT/BB*, Railinfrabeheer, Utrecht.

Projectteam Samensporen (2006) *Dynamisch Verkeersmanagement-proef Schiphol 2005*, Utrecht.

ProRail (2003) *Handboek treindienstleider*, ProRail, Utrecht.

ProRail (2004) HPO in de praktijk, in: *ProRail Special - Magazine voor medewerkers* (feb): pp. 6-7.

ProRail (2005a) *Algemeen Voorschrift C5515/I*, ProRail, Utrecht.

ProRail (2005b) *Beheerplan 2005*, ProRail, Utrecht.

ProRail (2006) *Veiligheidsmanagementsysteem ProRail*, ProRail, Utrecht.

ProRail (2007) *Jaarbericht 2006*, ProRail, Utrecht.

ProRail (2008a) *Afsluitende BB21 Rapportage*, ProRail, Utrecht.

ProRail (2008b) *Financieel jaarverslag 2007*, ProRail, Utrecht.

ProRail (2009a) *Beheerplan 2009*, ProRail, Utrecht.

ProRail (2009b) *Netverklaring 2010*, ProRail, Utrecht.

ProRail, Nederlandse Spoorwegen and Belangenvereniging rail goederenvervoerders (2008) *Programma Hoogfrequent Spoorvervoer - Verkenning benodigd investeringsvolume*, Utrecht.

Raad voor de Transportveiligheid (2001) *Botsing tussen twee reizigerstreinen in Dordrecht 28 november 1999*, Raad voor de Transportveiligheid, Den Haag.

Railforum Nederland (2001) *GroeiSporen - Alternatieve dienstregelingsmodellen en kleinschalige aanpassingen aan de infrastructuur ter vergroting van de vervoerscapaciteit van het spoor - Hoofdrapport*, Railforum, Hoofddorp.

Railned (2002) *Seinenboek*, Railned, Utrecht.

Railned Spoorwegveiligheid (2002) *Trendanalyse 2001 - trends in veiligheid van het spoorwegsysteem in Nederland*, Railned, Utrecht.

Railway Gazette International (04 January 2008) ETCS software error led to Lötschberg derailment

Rasmussen, J. (1983) Skills, Rules and Knowledge; Signals, Signs and Symbols, and other distinctions in human performance models, in: *IEEE transactions on systems, man, and cybernetics* SMC-13(3): pp. 257-266.

Rasmussen, J. (1997) Risk management in a dynamic society: a modelling problem, in: *Safety Science* 27(2-3): pp. 183-213.

Reason, J. (1990) *Human error*, Cambridge University Press, Cambridge, UK.

Reason, J. (2000) Safety paradoxes and safety culture, in: *Injury control & safety promotion* 7(1): pp. 3-14.

Reason, J., E. Hollnagel and J. Peries (2005) *Revisiting the "Swiss Cheese" Model of accidents*, Eurocontrol, Brétigny-sur-Orge, Cedex, France.

Regeling Keuring Spoorvoertuigen (2004), 's Gravenhage.

Regeling Spoorverkeer (2004), 's Gravenhage.

Regeling veiligheidsattest hoofdspoorwegen (2004), 's Gravenhage.

Regtuijt, P. (2007) *Organisatie van het werk van treindienstleiders - zelfsturende teams: een mogelijkheid tot meer variatie in het werk?*, Rijksuniversiteit Groningen & ProRail, Groningen.

Rensink, R. A., J. K. O'Regan and J. J. Clark (1997) To see or not to see: The need for attention to perceive changes in scenes, in: *Psychological Science* 8(5): pp. 368-373.

Rietveld, P. (2005) Six reasons why supply-oriented indicators systematically overestimate service quality in public transport, in: *Transport Reviews* 25(3): pp. 319-328.

Rolt, L. T. C. (1966) *Red for danger*, David & Charles, Newton Abbot, Devon.

Roth, E. M., N. Malsch, J. Multer and M. Coplen (1999) *Understanding How Train Dispatchers Manage and Control Trains: a cognitive task analysis of a distributed team planning task*, Proceedings of the human factors and ergonomics society 43rd annual meeting., Houston, Texas, pp. 218-222.

Roth, E. M., J. Multer and T. Raslear (2006) Shared situation awareness as a contributor to High Reliability Performance in railroad operations, in: *Organization Studies* 27(7): pp. 967-987.

Roth, G. (2006) Gleichtakt im Neuronennetz, in: *Gehirn&Geist* (1): pp. 24-32.

Rüter, A. J. C. (1960) *Rijden en staken - de Nederlandse Spoorwegen in oorlogstijd 1940 - 1945*, Martinus Nijhoff, 's Gravenhage.

Saari, J. (1986) Accident epidemiology, in: M. Karvonen and M. I. Mikheev (eds.) *Epidemiology of occupational health,* World Health Organization, Copenhagen.

Saarloos, A. (2008) *Risicoanalyse matrixsignaalgevers met beelden gevormd door geel licht*, ProRail, Utrecht.

Sanders, F. (2005) *STS-reductie - bijdrage Prorail Verkeersleiding*, ProRail, Utrecht.

Schaafsma, A. A. M. (2001) *Dynamisch Railverkeersmanagement*, Delft University Press, Delft.

Schaafsma, A. A. M. (2004) *Innovatieve oplossingen voor spoorknelpunt Schiphol van studie naar implementatie*, Colloquium Vervoerplanologisch Speurwerk 2004, Zeist.

Schaafsma, A. A. M. and V. A. Weeda (2009) *Operation-driven scheduling approach for fast, frequent and reliable railway services*, International Seminar on Railway Operations Research / RailZurich2009, Zurich.

Schaafsma, A. A. M. and E. A. G. Weits (1996) *Naslagwerk capaciteit en belasting van het spoorwegnet*, RailNed, Utrecht.

Schaal, V. (2007) *Energy-efficient driving - A change management programme of Deutsche Bahn's Passenger Transport Division*, International Transport Forum, Workshop on Ecodriving, Paris.

Schmeink, B. and M. Beuk (2006) *Voortgangsrapportage Passages Stoptonend Sein (STS) - Spoorbranche - stand van zaken eind 2005*, Nederlandse Spoorwegen, ProRail, Railion, Utrecht.

Schoemaker, T. J. H., K. Koolstra and P. H. L. Bovy (1999) Traffic in the 21th century - a scenario analysis for the traffic market in 2030, in: M. P. C. Weijnen and E. F. ten Heuvelhof (eds.) *The infrastructure playing field in 2030,* Delft University Press, Delft, pp. 175-194.

Schotanus, B. (2004) De dienstregelingbug, in: *OV-Magazine* april 2004.

Schotanus, B. and L. Zigterman (2005) *De orde van het seinhuis II - verslag van onze ervaringen*, Colloquium vervoerplanologisch speurwerk, Antwerpen.

Schultz van Haegen, M. H. (4 July 2003) *Kamervragen*, 's Gravenhage, pp. DGP/SPO/U.03.01726.

Schwanhäußer, W. (1974) *Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn*, RWTH Aachen, Aachen.

Schwarz, S., M. Hammerl and F. Feldmann (2009) Quantifizierung menschlicher Fehler für Risikoanalysen, in: *Signal + Draht* 101(6): pp. 19-23.

Schweizerische Fahrdienstvorschriften FDV (R 300.1–.15) (2006), BBL, Bern.

Sheridan, T. B. (2004) Driver Distraction From a Control Theory Perspective, in: *Human Factors* 46(4): pp. 587-599.

Sierts, A. C. F. (2007) *Het spoor van Dalen naar succes fase II - Economie van klantwaarde op nummer 1 - genomineerde inzending ProRail Priisvraag*, Voorburg.

Sierts, A. C. F., W. Wiersema and T. Lindhout (2007) *Eco - nomisch spoorverkeer met de integrale groene golf*, Utrecht.

Simon, H. A. (1957) *Models of man: social and rational; mathematical essays on rational human behaviour in a social setting*, Wiley, New York.

Simons, D. J. and C. F. Chabris (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events, in: *Perception* 28(1059-1074).

Sivak, M. (2002) How common sense fails us on the road: contribution of bounded rationality to the annual worldwide toll of one million traffic fatalities, in: *Transportation Research Part F* (5): pp. 259-269.

Slob, Van Hijum, Hofstra and Verdaas (2000) *Motie van het lid Slob c.s.*, Tweede Kamer der Staten-Generaal, 's Gravenhage.

Snow, J. (1854) *On the mode of communication of cholera*, 2nd edition by Delta Omega, Piccalilly.

Spoorwegongevallenraad (1993) *Botsing tussen intercity 846 en stoptrein 15243 te Eindhoven op 31 oktober 1992*, Spoorwegongevallenraad, 's Gravenhage.

Spring, P., A. McIntosh, C. Caponecchia and M. Baysari (2009) *Level of Automation: effects on train driver vigilance*, 3rd international rail human factors conference, Lille.

Stamatelatos, M. and W. Vesely (2002) *Fault tree handbook with aerospace applications*, NASA, Washington DC.

Starbuck, W. H. and F. J. Milliken (1988) Challenger: fine-tuning the odds until something breaks, in: *Journal of Management Studies* (25): pp. 319-340.

Steenhuisen, B. and M. De Bruijne (2009) *The brittleness of unbundled train systems: crumbling operational coping strategies*, Second International Symposium on Engineering Systems, MIT, Cambridge, Massachusettes, MIT ESD.

Stolk, A., F. Makkinga and J. Casteel (2000) *Memo - verslag van de tussenresultaten CS/BO proef op Utrecht IV*, Holland Railconsult - unpublished document, Utrecht.

Strauss, A. and J. Corbin (1998) *Basics of Qualitative Research - techniques and procedures for developing grounded theory*, Sage Publications, London.

Studiecommissie ATB (5 juli 1963) *Verslag nr. 25*, Nederlandsche Spoorwegen, Utrecht.

Studiecommissie ATB (22 november 1963) *Verslag nr. 39*, Nederlandsche Spoorwegen, Utrecht.

Studiecommissie ATB (23 januari 1964) *Verslag nr. 45*, Nederlandsche Spoorwegen, Utrecht.

Studiecommissie ATB (26 juli 1963) *Verslag nr. 28*, Nederlandsche Spoorwegen, Utrecht.

Studiecommissie ATB (26 oktober 1967) *De overwegingen van de voorstanders van een "licht" remkriterium*, Nederlandsche Spoorwegen, Utrecht.

Swuste, P. H. J. J., C. van Gulijk and A. W. Zwaard (2008) *Human Behavior: The first safety theory. Accident causality in the 19th and the start of the 20th century, the rise of the accident proneness theory in Great-Britain, the United States and its consequences in The Netherlands* Working on Safety 2008, Crete, Greece.

Theeg, G. and B. Vincze (2007) European train protection systems compared, in: *RTR* (4): pp. 10-16.

Thompson, L. S. (2001) *World bank's railways database*, World Bank, Washington D.C.

Transrail (2007) *CATO System Requirements Specification*, Transrail Sweden AB, Sundbyberg.

Trouw (4 aug 2008) Prijs treinkaartje schiet omhoog.

Tweede Kamer der Staten-Generaal (2005a) *Veiligheid van het railvervoer - Lijst van vragen en antwoorden (vastgesteld 16 december 2005)*, SDU Uitgevers, 's Gravenhage.

Tweede Kamer der Staten-Generaal (2005b) *Vragen van het lid Gerkens (SP) aan de minister van Verkeer en Waterstaat over spoorveiligheid. (Ingezonden 17 januari 2005)*, SDU Uitgevers, 's Gravenhage.

Tweede Kamer der Staten-Generaal (2007) *Veiligheid van het railvervoer - lijst van vragen en antwoorden*, SDU Uitgevers, 's Gravenhage.

Tweede Kamer der Staten-Generaal (2008a) *Spoor: vervoer- en beheerplan - brief van de Minister van Verkeer en Waterstaat*, SDU Uitgevers, 's Gravenhage.

Tweede Kamer der Staten-Generaal (2008b) *Vragen van het lid Roemer (SP) aan de minister van Verkeer en Waterstaat over spoorveiligheid. (Ingezonden 11 juni 2008)*, SDU Uitgevers, 's Gravenhage.

UIC (1998) *ERTMS (ETCS/EIRENE) MMI*, UIC, Paris.

UIC (2004) *Code 406 - Capacity*, UIC, Paris.

UIC (2008) *Leaflet 612 - Display System in Driver's Cabs (DDS) - General Requirements, Set-Up and Technical Specifications (preliminary version)*, UIC, Paris.

Vaessens, B., M. van Hagen and M. Exel (2008) *Auto vs trein - de context bepaalt de keuze*, NS Commercie, Utrecht.

van Alphen, W. J. T., J. Gort, K. I. J. Stavast and A. W. Zwaard (2008) *Leren van ongevallen - een overzicht van analysemethodieken*, SDU Uitgevers, Den Haag.

van Bruxvoort, M., H. Koot, E. van de Loo and E. Steenhuisen (2009) *Marktobservatie Rail 2008*, TNS Consult, Amsterdam.

Van den Boom, A. J. J. and B. De Schutter (2007) *On a model predictive control algorithm for dynamic railway network management.*, 2nd International Seminar on Railway Operations Modelling and Analysis (RailHannover2007), Hannover.

van den Hout, J. (1992) Vervallen van de ATB-kwiteerfunctie, in: *Matblad* 1992(2): pp. 29-31.

Van den Top, J. (2005) *Toepassingsmogelijkheden van dynamisch railverkeersmanagement in Arnhem*, TU Delft / ProRail, Delft.

Van den Top, J. (2006) *Effectiveness of performance indicators and process control in Dutch rail transport*, 9th TRAIL Congress 2006, TRAIL in MOTION, Rotterdam, TRAIL Research School, pp. 1-17 (CD-ROM).

Van den Top, J. (2009) *Ongeldige onderbouwing bij wijziging treinbeveiligingssysteem - de ATB-kwiteerfunctie: irriterend of intrigerend?*, NVVK Veiligheidscongres, Papendal, Nederlandse Vereniging voor Veiligheidskunde.

Van den Top, J. and J. Groeneweg (2007) IJzeren regels geen waarborg voor spoorwegveiligheid, in: *Tijdschrift voor Toegepaste Arbowetenschap* 20(3+4): pp. 23-27.

Van den Top, J. and A. C. F. Sierts (2009a) ERTMS in Nederland, in: *Op de Rails* (3): pp. 133-140.

Van den Top, J. and A. C. F. Sierts (2009b) Railverkeer veilig en economisch regelen, in: *Op de Rails* (1): pp. 23-32.

Van den Top, J. and B. Steenhuisen (2009) Understanding ambiguously structured rail traffic control practices, in: *International Journal of Technology, Policy and Management* 9(2): pp. 148-161.

Van der Weide, R., H. Frieling and D. de Bruijn (2009) *(Why) are dutch cargo trains 2.6 times more often involved in spads compared to passenger trains?*, 3rd international rail human factors conference, Lille.

Van Dijk (26-6-1920) *KIvI-lezing*.

van Eeten, M. J. G., E. Roe, P. R. Schulman and M. d. Bruijne (2006) When failure is not an option, in: R. M. Verburg, J. R. Ortt and W. M. Dicke (eds.) *Managing Technology and Innovation, An Introduction,* Routledge, Oxon, pp. 306-321.

Van Gijzel, Weekers, Eurlings, Stellingwerf and Van der Steenhoven (2000) *Motie van het lid Van Gijzel c.s.*, Tweede Kamer der Staten-Generaal, 's Gravenhage.

van Luipen, J. and S. Meijer (2009) *Uploading to the MATRICS: Combining simulation and serious gaming in railway simulators*, 3rd international rail human factors conference, Lille.

van Westrenen, F. (1999) *The maritime pilot at work - evaluation and use of a time-to-boundary model of mental workload in human-machine systems*, TRAIL, Delft.

Veeneman, W. W. (2002) *Mind the gap - bridging theories and practice for the organisation of metropolitan public transport*, TRAIL, Delft.

Veenman, A. W. (12 september 2006) *Dienstregeling 2007*, Nederlandse Spoorwegen, Utrecht.

Verein Deutscher Ingenieure (2007) *Sicherung von Anlagen der Verfahrenstechnik mit Mitteln der Prozessleittechnik (PLT) - Einführung, Begriffe, Konzeption*, Beuth Verlag, Berlin.

Verheij, F. and R. P. van Dijk (2008) *Voortgangsrapportage Passages Stoptonend Sein (STS) - Spoorbranche - stand van zaken 31-05-2008*, Spoorbranche, Utrecht.

Vervoerconcessie hoofdrailnet (2005), Den Haag.

Vicente, K. J. (1999) *Cognitive Work Analysis: towards safe, productive and healthy computer based work*, Lawrence Erlbaum, Marwah, NJ.

Vieira, G. E., J. W. Herrmann and E. Lin (2003) Rescheduling Manufacturing Systems: A Framework of Strategies, Policies, and Methods, in: *Journal of scheduling* 6: pp. 39-62.

Visser, J. P. (1998) Developments in HSE management in oil and gas exploration and production, in: A. R. Hale and M. Baram (eds.) *Safety Management: the Challenge of change,* Pergamon, Oxford, pp. 43-65.

von Foerster, H. (1985) *Sicht und Einsicht: Versuche zu einer operativen Erkenntnistheorie*, Friedrich Vieweg und Sohn, Braunschweig.

von Meier, A. (1999) Running in the other team's shoes: different occupational cultures challenge what defines technological innovations, in: *IEEE Potentials* 18(2): pp. 21-25.

Vorage, P. (2006) *De effecten van planning en bijsturing op capaciteit en punctualiteit*, TU Delft, Delft.

Vromans, M. J. C. M. (2005) *Reliability of Railway Systems*, Trail research school, Delft.

Walgemoed, H. (2005) Flight envelope, in: F. N. Stoliker (ed.) *Introduction to Flight Test Engineering,* Research and Technology Organisation, North Atlantic Treaty Organisation, Neuilly-sur-Seine Cedex, France.

Wedzinga, A. A. (2002) *Seinenboek - toelichting*, Railned, Utrecht.

Weeda, C. E. (1999) *"Vergissen is menselijk" - STS en menselijke fouten*, AMG\Ergonomie, Utrecht.

Weeda, V. A. (2005) *Plannen met speling, spelen met de planning - Naar een betrouwbare treindienst door terugkoppeling van uitvoering naar planning*, TU Delft, Delft.

Weeda, V. A. and K. S. Hofstra (2008) *Performance analysis: improving the Dutch railway service*, Computers in Railways XI, Toledo, WIT Press.

Weeda, V. A. and P. B. L. Wiggenraad (2005) *Betrouwbare treindienst door terugkoppeling van uitvoering naar planning*, Verkeerskundige Werkdagen 2005, Ede, CROW.

Weeda, V. A., P. B. L. Wiggenraad and K. S. Hofstra (2006) *Een treinvertraging zit in een klein hoekje. Resultaten punctualiteitsanalyse casestudy Rotterdam - Dordrecht*, Colloquium Vervoersplanologisch Speurwerk (CVS), Amsterdam, CVS, pp. 445-456.

Weick (et. al.) (2002) High Reliability - The power of mindfulness, in: F. Hesselbein and B. R. Johnson (eds.) *On high performance organizations - a leader to leader guide,* The Drucker Foundation, Jossey-Bass, San Francisco, pp. 7-18.

Weick, K. E. and K. M. Sutcliffe (2001) *Managing the Unexpected: Assuring High Performance in an Age of Complexity*, Jossey Bass, San Francisco.

Weits, E. A. G. (1997) *BB21-verkeersregeling - vensters geven uitzicht op verkeersregeling*, Railned, Utrecht.

Wickens, C. D. and J. G. Hollands (2000) *Engineering psychology and human performance*, Prentice Hall, Upper Saddle River, New Jersey.

Wilde, G. J. S. (1994) *Target Risk*, PDE Publications, Toronto, Canada.

Wilson, J. R., T. Farrington-Darby, G. Cox, R. Bye and G. R. J. Hockey (2007a) The railway as a socio-technical system: human factors at the heart of successful rail engineering, in: *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* (1): pp. 101-115.

Wilson, J. R., B. Norris, T. Clarke and A. Mills (2005) *Rail human factors: supporting the integrated railway*, Ashgate, Aldershot.

Wilson, J. R., B. Norris, T. Clarke and A. Mills (2007b) *People and rail systems: human factors at the heart of the railway*, Ashgate, Aldershot.

Wilson, J. R. and B. J. Norris (2005) Rail human factors: past, present and future, in: *Applied ergonomics* 36: pp. 649-660.

Winter, P. (2009) *Compendium on ERTMS*, Eurailpress, Hamburg.

Woods, D. D. and E. Hollnagel (2006) *Joint cognitive systems*, Taylor & Francis, New York.

Wright, L. and J. R. Vorderegger (2005) *Selectiemethodiek risico van seinen - beschrijving en validatie*, STS-werkgroep (ProRail; Inspectie Verkeer en Waterstaat), Utrecht.

Wüst, R., F. Laube, S. Roos and G. Caimi (2008) *Sustainable Global Service Intention as objective for Controlling Railway Network Operations in Real Time*, 8th World Congress of Railway Research (WCRR), Seoul, Korea.

Young, M. S. and D. P. J. Grenier (2009) *Future train cab interface design: development of a model to support driver situation awareness*, 3rd international rail human factors conference, Lille.

Zigterman, L. (1985) NS op weg naar beheersing treinverkeer, in: *Technovisie* (february): pp. 10-14.

Zoeteman, A., M. Deerenberg, S. Heijstek and A. Schaafsma (2009) Facilitating traffic growth in the Netherlands: the 'Triple A' strategy approach, in: *European railway review* (1): pp. 14-19.

PART V: Appendices

Appendix A Examples of current projects

In this appendix, several examples of ongoing projects in the rail sector will be given. Since nearly all of these projects are still ideas or only recently applied, it is not yet possible to draw lessons from their application. Still, it is worthwhile to describe them, since the objectives aimed at in the projects fit to a great extent into the theoretical framework developed in this thesis, but by using this framework, modifications or extensions can be proposed to make them more effective or more coherent.

A1 Customer-oriented logistics
ProRail realises that several train operating companies have different goals they are striving for. This means that they also have different requirements for the delivery of train paths. A passenger train operating company may give a high priority for having their trains running to schedule to ensure connections, where a freight train company might rather have train paths in which their trains do not have to stop so that they can save on fuel. In an ongoing project that started in 2008, ProRail aims to translate such different requirements into terms directly relevant to local and regional traffic controllers. This means that the functions of slot design, slot allocation and slot use will be explicitly addressed. Every train path needs to be fitted into a slot. Thus, the slots serve as an explicit tolerance for rescheduling and make the traffic pattern predictable. Also, the network will be divided into congestion zones and buffer zones; in congestion zones, the main objective is to keep trains moving in order to avoid cluttering the bottlenecks. As a result, trains may come out of the bottleneck in another pattern than what is desired for their later journey. In buffer zones, the main objective is to 'repair' such delays and wrong orders between trains, and provide the ability to put trains into the (next) bottleneck in the right order and at the right moment. This means that buffer zones need a track layout that allows for such actions, such as waiting (=consuming superfluous buffer time when the train did not meet a disturbance which has been allowed for) or overtaking, but also to remove trains from the system or insert them into it.

The division of the network into buffer zones and bottlenecks can help designers of infrastructure, timetable and real time traffic control to focus on specific control measures for such areas. For instance it should be ensured that trains can leave the bottleneck as quickly as possible. The clear division in the tasks of slot design, slot allocation and slot use fits well onto the functions mentioned in Table 10.1. In the light of this research, this approach seems to be a good idea; the several requirements could be translated into operational envelopes and as such indicate which variables are important to monitor. These (and the adherence of the process to the required values) should then be made visible to the controllers.

A2 New design norms
Currently, ProRail is developing a new set of infrastructure design norms. It starts with the necessary functions, with functional requirements. From these, the requirements are gradually made more detailed and specific in terms of physical system elements. The following order is being introduced:
- Functions: e.g. traction energy supply, traffic management, train protection, etc.
- Processes: for the function of train protection e.g. train runs, shunting, maintenance and construction, etc.
- Conceptual solutions: for the process of train runs e.g. lineside signalling, cab signalling, automatic train protection, etc.
- Physical products: e.g. interlocking, signal, switch controller, etc.

The approach taken broadly fits the recommendations of this research, which is that processes should be designed first and only afterwards should it be decided what the requirements are for technology to support these processes. It may be beneficial to define the functions on the infrastructure somewhat more precisely than just 'driving a train', but relate them to the driving regimes as they were introduced in section 10.8.3 or to similar standard solutions to be taken by traffic control (e.g. overtaking). These regimes, and therefore also their application in design norms, requires further research.

A3 Operational Control Centre Rail (OCCR)
The OCCR (Operational Control Centre Rail) was installed in reaction to a major collapse of the traffic management systems in Utrecht due to a technical failure, so that the rail traffic in large parts of the Netherlands was disturbed for almost two days (ProRail 2007). It is meant to combat such severe disruptions more effectively. The OCCR provides a control room where several organisations, such as ProRail, train operating companies and, for instance, public emergency services, can be present so that they can all have access to the same information, which can be displayed on video walls. This is meant to shorten communication lines between these actors and improve their mutual understanding of the situation. The OCCR is running as a trial for several years, and has already shown clear benefits in solving service disruptions more quickly, although it is not yet a formal part of the organisation's control model.

A4 (De)centralised traffic control
Another initiative, and one desired by many traffic controllers who were interviewed for this thesis, is a recent initiative to change the structure of their traffic control department as it was described in sections 2.3.6 and 2.3.7. This is meant to improve the quality of service and also to be able to differentiate between different customers' preferences, as was described under the heading of customer-oriented logistics. The local traffic control task will remain the same, with seventy-five traffic controllers based in twelve control centres. The four regional traffic control centres will be discontinued and the control organisation will then look as follows:
- one central, nationwide control centre and
- twelve control centres, each with
  - one decentralised (regional) controller and
  - several local traffic controllers

The new structure aims to provide quicker responses to disturbances and will have been put into service by the end of 2009. One objective is to avoid suboptimisation which occurred when focussing on four regions instead of the whole network. Another improvement will be the communication between local and regional traffic controllers, which will become possible by shorter reaction times and better mutual understanding of why decisions are being made. This will become possible, since the decentralised regional controller will now be located in the same traffic control rooms as the local controllers. The decentralised regional controller will reschedule trains within the area he is responsible for in close cooperation with the local traffic controllers. The two staff members that make up the central network control centre will act as supervisors, for instance to monitor the progress of 'problematic trains', to decide what to do in cases where several disturbances happen simultaneously, or to schedule ad-hoc requests for extra trains. Finally, a director will be in place to overcome impasses and to provide feedback for learning. The central traffic control centre will also be part of the OCCR. Although the two projects started separately, they form a logical combination. Several trials were performed which showed clear benefits in the co-operation; it is expected that things can improve even more when the trial situation is made formal, since also the staff education can then be made specific to the new situation.

During interviews for this thesis, it was indicated that at some traffic control centres this working method had already evolved by itself. According to the theoretical framework, shortening of communication times and sharing the same situation awareness by placing operators in the same control room (either for OCCR or for decentralised controllers in a traffic control centre) is basically a good idea.

A5 Dynamic Traffic Management (DTM)
A recent development for local traffic controllers is that of Dynamic Traffic Management[18]. DTM started as a notion that dynamic events do happen in rail traffic and offered some ideas on how these events could be controlled more effectively. Originally, several innovative ideas were grouped under this heading, such as improved traffic management systems, advisory speeds for train drivers, and scheduling differently to allow more freedom for operators to solve conflicts at short notice – the original aims are somewhat like those formulated in the project 'customer-oriented logistics', but they were production-driven rather than customer-driven.

The ambitious objectives of DTM have been abandoned when it came to implementing these ideas, for the following reasons:
1. although dynamic traffic management suggests that it is 'new and better', no clear comparison to the current situation was possible (static traffic management?) so that it was difficult to convince the decision makers that a new approach would be indeed better.
2. dynamic traffic management offers a high number of control measures that could be taken, but it was not clear how these would interact or which measure should be used in which case. ProRail started to explain dynamic traffic management as a 'container full of measures' (in Dutch both with the acronym DVM), but has failed as yet to develop a clear picture of their relationship and their potential for achieving higher quality services.
3. no clear goal regarding the allowable dynamics in the traffic process has been formulated to which dynamic traffic management – or its static counterpart – should contribute. This goal has now hopefully been found in the KPI 'train path' (section 2.2.5) and/or in 'customer-oriented logistics'.

Finally, a much less sophisticated, yet effective application of some DTM measures has been applied to railway tunnel under Schiphol airport. In order to allow a higher number of trains to pass through the tunnel, several measures have now been implemented:
- The timetable has been changed such that no running time margins are included on the tunnel stretch, so that trains do not have to wait within the bottleneck. Outside the tunnel, the running time margin is given back in order to recover from possible delays and to allow trains to continue their journey according to schedule.
- Small conflicts are allowed to be contained in the timetable and will be solved during operations on a first come, first served basis.
- The platform side to which trains are assigned is alternated, to increase the probability that trains can enter a free platform track.

These measures allow traffic controllers to focus more on the overall picture and less on the details such as train orders or re-routings. Also, the example shows that if timetabling and operations are considered simultaneously, more trains can be scheduled, which again illustrates the need to take all functions in the cascade model explicitly into account. After a trial, railway traffic at Schiphol is now normally treated in this way.

[18] Note that the word choice "dynamic traffic management" is in fact incorrect, since not the management itself is dynamic, but the dynamics in the traffic process are managed.

A6 European Train Control System

The European Train Control System (ETCS) is a European standardised set of interfaces for train control. This was deemed necessary by the European Union to overcome increasing problems with international train traffic, since every country had developed their own incompatible train protection systems. This section will focus on ETCS's capability to support the driver's process; a broad overview of ETCS and its accompanying developments is given elsewhere in English (Winter 2009) or in Dutch (Van den Top and Sierts 2009b; a). The European Train Control System (ETCS) is frequently explained in terms of Levels. Every Level defines a different technical layout for how data is sent to and from the train. However, for the purpose of this thesis the distinction between Levels is not so relevant, since all 'genuine' ETCS Levels (1, 2 or 3) provide the same functionality (Winter 2009), being:

- Standardised information exchange between trackside and train;
- Standardised information exchange between train and driver (DMI: Driver Machine Interface, mainly cab signalling).
- Train supervision (automatic train protection).

The functionality of ETCS is defined in so-called (software) baselines. At the time of writing, the current version is 2.3.0d. As of 2012, baseline 3.0.0 will arrive with many new features and problems of the current version solved, such as the very conservative braking model. Currently, the DMI is a non-mandatory interface so that small differences between DMIs from several suppliers exist (see Figure A1 for an example of one of the first versions). This will also be harmonised when the DMI becomes a mandatory standard as of baseline 3.0.0. The current requirements for the DMI can be found in the specifications by the European Railway Agency (2009).

ETCS 'modes' define different phases of the driving process and allocate responsibilities between the on-board computer and the driver. Normally when people refer to ETCS, they implicitly refer to the Full Supervision mode, in which the driver controls the train and ETCS fully supervises whether he does so within the limitations of the movement authority. Other modes are for example SHunting, On Sight, Staff Responsible, in which staff has somewhat more freedom and responsibilities. In the TRip mode, the on-board computer intervenes and takes over control by applying an emergency stop.

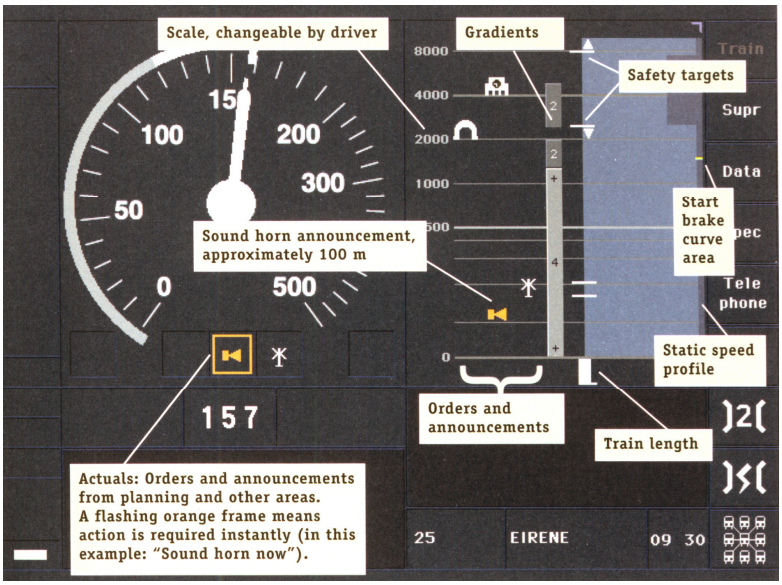


Figure A1: Early design for the ETCS Driver Machine Interface (UIC 1998; reproduced with permission).

ETCS can be helpful to improve at least the following of the functions mentioned in Table 10.1:

- *Train control on path*: ETCS does not currently provide any information about the (re)scheduled path. However, it could do so and this was intended in earlier stages of the system's design. There is a part of the ETCS DMI called the Planning Area, which provides information for the driver to anticipate events. This is a step forward, but currently, it shows only infrastructural objects and limitations, and no information on the traffic controller's plan or the train's path adherence, so that it does not fulfil the requirements of this thesis. A further step was originally catered for, since the ETCS DMI reserved room to display an advisory speed. This can serve as a rudimentary indication of what must be done to achieve the goal state (without being explicit about what this goal is) and would help the driver to follow his path closely. However, making use of this functionality is not mandatory for companies using ETCS. At the moment, ETCS is strictly speaking (despite its name) not a train control system, because it does not make sure that the train behaves according to the plan of the traffic control centre, but several proposals have been made to display more detailed traffic management information such as the intended path that the driver should follow in the ETCS Planning Area (Sierts et al. 2007; Transrail 2007).
- *Signalling*: ETCS can calculate the Most Relevant Speed Profile (MRSP; the minimum of all restrictions e.g. due to infrastructure design, route setting, temporary speed restrictions, maximum train speed, etc), thus taking this task over from the driver. It calculates the braking distance, thus showing a 'mobile distant signal' on the Planning Area at the position where the train must start to brake. This would operationalise an important aspect of the controllability envelope of the safe envelope model. ETCS cab signalling provides much more information than lineside signalling. The problems described in Chapters 7 and 9 would be solved, since the variety of the information is sufficient. Also, visual problems with lineside signalling such as sighting distances or relating the signal to the track will be solved when cab signalling is used.
- *Train protection*: Not only can ETCS check whether the driver remains within the braking curve that fits the MRSP, it can also check other dimensions. Under the heading of route suitability, it can check loading gauge, power supply, or axle load. Although such issues did not receive much attention during this thesis, these are also part of the system state and the safe envelope principles could be applied to them, too.
- *Traffic control:* ETCS, as a standardised information exchange system, could also be used for traffic management functions and such functions are foreseen. Experiments that fit into this chapter are Combine, Combine II (Mazzarello and Ottaviani 2005) and Europtirails (Winter 2009). Many railways currently see the need for some kind of traffic management system. ETCS may play a useful role in providing (parts of) the technical infrastructure for traffic management functions, such as a driver display, a communication system (GSM-R) and an on-board computer which knows most of the train's status information (e.g. speed, position, length). This will make it relatively easy to get information to and – equally important – from the trains, and is practiced on a small but effective scale in the Swiss Lötschberg base tunnel (Montigel 2008). Using ETCS to exchange data may also have a standardisation advantage compared to several national solutions currently being in development.

In conclusion, the ETCS DMI fits well with the requirements from the safe envelope framework, yet it only focuses on not exceeding the operational envelope and not on steering the process within the envelope. It is, however, technically easily feasible to add the latter functionality.

A7 RouteLint

RouteLint (English: RouteRibbon) is the product of an earlier project called 'Het Spoor Meester' (Master of the Rail). This project originally had a wider scope, which was to make traffic controllers and drivers work together much more closely, but this scope was gradually narrowed. RouteLint now provides information about the current traffic situation to drivers. The information which is shown was already available in the VPT system and is now also shown to the driver. It serves to display probable disturbances to the driving process in advance so that the driver can adopt a more anticipatory driving style. The information content is primarily the result of what was technologically easily available, and not the result of what the driver really needs to perform his processes. However, based on the initial design, some changes have been made to make the information better fit the driver's needs. The interface is shown in Figure A2. The following information is provided:

- Own train: position, adherence to current plan (in whole minutes), planned route and route setting with changes made by the traffic controller;
- Trains in front: positions, delays, route setting, whether these trains are entering, crossing or leaving the own train's route;
- Train behind: indicates which train will be hindered first if you were to stop now, and the location of that train.

RouteLint generally increases driver motivation. In a field trial drivers managed to encounter fewer signals at 'warning' and, as a consequence, there was less need to brake. This resulted in benefits such as punctuality improvements (from 86% to 87.5%) and energy savings of approximately 3-5%, especially for intercity services and freight trains. (Albrecht et al. 2007). In order to minimise the required time to overtake a local train, another trial was performed, but with another DMI. The distance to the train ahead and its speed were presented to drivers. Without encountering more restrictive signals, they could reliably reduce the overtaking time from six to four minutes, which increases capacity on the line (van Luipen and Meijer 2009).

RouteLint does not fully satisfy the requirements from the theoretical framework, but it is a step in the right direction. With RouteLint, the driver will have information about the situation ahead, but he has no clear goal towards which he should control, nor does he have a tolerance within which he should operate. However, he can now easily obtain information for which formerly he needed to contact the traffic controller by phone. Also, the information given to the driver is much richer than that provided by the signalling system, so that the driving process can benefit to a certain extent from the information shown. Although the information shown enables the driver to see what the traffic controller has done, he still cannot see what the traffic controller's plan is. Thus, the driver can anticipate only to a certain extent.

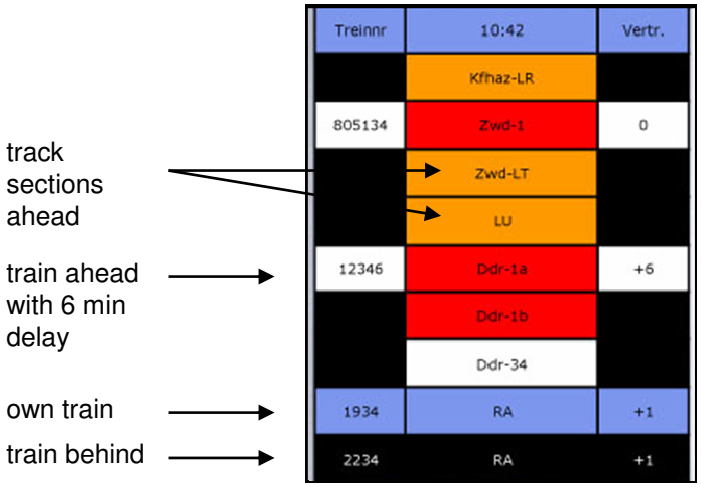


Figure A2: Interface of RouteLint (Source: ProRail).

Important missing items, that ProRail and Dutch Railways will try to solve in a new version of RouteLint, are, amongst many other planned new features:
- More frequent updates of delay information.
- Delay information in fractions of minutes.

A next version RouteLint will possibly be linked to support the driving process as defined in the project 'customer oriented logistics' (see above).

To ensure safety it is very important that the driver realises when he is approaching the controllability boundary. In order not to distract him from that, the RouteLint screen is automatically switched off when a signal is passed that is showing a yellow aspect ('prepare to stop'). This is a clear indication to the driver that he needs to focus fully on the immediate situation outside. The driver does not have to generate an intelligent driving strategy at this point as there is only one appropriate control measure: he needs to stop before the red signal. However, once the train comes to a standstill, the driver has the opportunity to temporarily override the black screen, and briefly assess the situation.
Care should be taken when driver assistance systems can distract the driver's attention from safety-critical information for which no automatic train protection exists: in the Dutch case for example in the approach to red signals or at temporary speed restrictions. For the latter, no solution has yet been found.

A8 Co-Production
The SBB Project PULS 90 includes under the title Co-Production, a control loop model for rail operations (Laube et al. 2007). The Co-Production starts from the idea that the process interface between the train operating company and the infrastructure provider is most simply described by two mutually agreed output promises:
1. The infrastructure provider provides a secured itinerary to the train operating company at an agreed time with an agreed maximum deviation (time window);
2. The train operating company promises in turn to make use of this itinerary within an agreed maximum deviation (time window).

As a consequence it is possible to model railway operation as a set of two control loops: an inner loop, where the executing actors of railway operations (driver, guard, signal operator) control their own process with the ambition to keep it within the agreed tolerance; and an outer control loop serves to coordinate their feedback (uplink) where they indicate process failure (inability to keep the process within the agreed tolerance) as input into developing new feasible process (time) targets - the rescheduling loop. This model allows the definition of simple common success criteria for all processes, that is absolute time and maximum allowable deviation. The expected core benefits of the model are in significantly reduced buffer time requirements between trains, which can be translated into additional trains or improved quality, depending on the challenge.

If a railway operates close to network saturation, high speed of reaction becomes pivotal to upholding production quality and efficiency. The control loop models show that the time at which a rescheduling decision is taken has a correlation with the amount of knock-on delays that can be prevented. The ability to accelerate trains to clear a track for a rescheduled train that will follow shortly afterwards, gives additional power to the rescheduling control loop, but it requires that this loop looks well ahead. Tests at SBB have shown that it is feasible to implement the double control loop approach and that precise control of train position and other operational processes is achievable within a 15 second tolerance (i.e. plus or minus 7.5 sec) if:

- The time criteria are achievable in the staff's perception;
- The staff have competences closely matched to their responsibilities (i.e. they are not asked to control processes which are beyond their control); and
- Staff have the ability to feed back process failures.

It turns out that the added precision and the better defined processes have a calming effect on the process owners and hence the technical systems they operate. SBB has made a decision in principle to evolve its operational processes to the Co-Production model.

We see that Co-Production fulfils most of the requirements for a train control system. In contrast to RouteLint, where information provided is mostly based on possible disturbances, the driver has no notion of these and is just presented with (the deviation from) a new path which was centrally defined and is conflict-free. This provides a clear goal with a defined tolerance. This motivates drivers and it allows the traffic control centre to monitor the progress of the process.

Appendix B Interview questionnaire

For interviews with drivers, local and regional traffic controllers, the following questionnaire was used.

Interviewprotocol

Geïnterviewde:

Datum:

Standplaats:

Interviewer:

Lengte interview:

Email:

Aantal jaar ervaring:

Vragen:
Niet los stellen, maar machinist gewoon laten vertellen wat hij/zij kwijt wil en zo af en toe checken welke vragen al beantwoord zijn.

!VRAAG ALTIJD NAAR HET WAAROM!

Antwoordformulier machinist n

1. Doelen (Goals)

De vragen gaan hier over wat u als machinist beoogt te bereiken. Wanneer hebt u een goede werkdag achter de rug?

1. Wanneer hebt u een goede werkdag achter de rug? Wanneer een slechte?
2. Hoe zou u het doel van uw werk omschrijven?

2. Huidige werkwijze (current practices)

In dit deel gaat erom dat u uw werkzaamheden omschrijft. Hoe bereikt u het doel dat u hierboven hebt geschetst? Wat is de afgesproken werkwijze? Kunt u daarmee het gestelde doel bereiken, of moet u er soms vanaf wijken? Zou u dat willen en kunnen?

Hebt u, of een collega, situaties meegemaakt die maar net goed zijn afgelopen? Waardoor ging het bijna fout en wat is er gebeurd waardoor de afloop toch goed is gegaan? Wat had je moeten weten zodat het niet fout zou gaan? Welke invloed heb je zelf op de veiligheid?

In geval van incidenten zult u vaak anders te werk gaan dan in het normale proces. Op welke manieren kunt u het proces tijdens een incident blijven besturen? Zijn daartoe aanvullende maatregelen nodig? Hoe gaat u om met werkzaamheden aan het spoor?

Operationele grens

3. Omschrijf uw werk. Hoe zou het moeten? Hoe gaat het in de praktijk?
4. Zijn er wel eens tegenstrijdige regels? In welke situaties? Wat zijn de gevolgen daarvan? Bent u bekend met alle regels (bv. ook if-then regels)?
5. Onder welke omstandigheden loopt de afgesproken werkwijze nu klem?
6. Ziet u soms betere oplossingen dan afgesproken in afhandelingsscenario's en de if-then-regels? Welke? Past u die dan ook toe? Zo nee, waarom niet?
7. Geef voorbeelden van situaties waarin iets niet loopt zoals u had gewild. Hoe komt dat? Is er sprake van strakke koppeling of complexiteit? Wat had je vooraf moeten weten om een betere* oplossing te kunnen kiezen ("Als ik dát had geweten....")? (beter qua veiligheid, tijdigheid van instellen, veiligheidscommunicatie, evt. ook treinpunctualiteit, energieverbruik, reizigersgemak, anders).

Besturingsgrens

8. Heeft u voldoende regelmogelijkheden om problemen op te lossen? Heeft u daar ook voldoende informatie voor?
9. Hoe veel vrijheid heeft u om af te wijken (HRO)? Wat doet u om het proces / de dienstverlening dan toch gaande te houden? Mag dat? Staat u daar ook bij stil? Wat zou volgens u wel moeten kunnen, maar mag niet? En wat mag wel, maar doet u niet (= niet opvullen envelop)?

Veilige grens

10. Wat verstaat u onder "veiligheidsrelevante communicatie"?
11. Hoe schat u in of iets zonder gevaar kan? (evt. voorbeelden: deelrijwegen, R.O.Z., tegen elkaar instellen van rijwegen, werkzaamheden, verzakkingen)
12. Welke handelingen die u moet verrichten vindt u onveilig? Wat zou er moeten veranderen om het veiliger te maken?

3. Informatievoorziening (Information supply)

Dit onderdeel behandelt uw informatiebehoefte over het proces dat u bestuurt. Welke informatie heeft u nodig, en waar komt die vandaan? Op welke zaken let u allemaal tijdens het rijden? Heeft u genoeg aan de seinbeelden of zou u soms meer willen weten (bv. plaats, afstand of spoor waar u het volgende sein kunt verwachten)? Krijgt u wel eens overbodige informatie of komt u tekort? Komt u wel eens voor verrassingen te staan?

Er wordt veel gesproken over het al dan niet gevaarlijk zijn van een verwachtingspatroon bij machinisten. Wat is uw ervaring hiermee en denkt u dat dit probleem groter of kleiner wordt wanneer het drukker wordt op het spoorwegnet? Welke informatie zou u nodig hebben om een eventueel onterechte verwachting te doorbreken?

13. Welke informatie zou je willen hebben? Wat krijg je? Wat ontbreekt er dus? Wat heb je dus te veel?
14. Welke informatie heeft u nodig om een goed beeld te kunnen vormen van de actuele situatie en goede beslissingen te kunnen nemen? Waar komt die informatie vandaan?
15. Welke aanvullende informatie komt daarbij indien er sprake is van verstoringen/vertragingen?

4. Toekomstige werkwijze (Ideas on future practices)

Stel dat er in de toekomst 2x zo veel treinen zouden moeten gaan rijden en er maar zeer beperkt budget is om dit mogelijk te maken. Er is wel tijd en geld voor aanpassingen in bedienings- en communicatiesystemen en aanpassingen van regels en organisatie.

Wat zijn sterke punten aan de huidige manier van werken die moeten worden behouden? Wat zou je zelf kunnen bijdragen aan een toename in het treinverkeer?

Wat zou er moeten worden veranderd om meer mogelijk te maken? Waarom? Welke mogelijkheden zijn er buiten de techniek om (organisatie/communicatie/informatie)?

16. Zijn er situaties waarin je slimmer bent dan de regels, m.a.w. zijn de regels te beknellend? Kun je meer dan je mag?
17. Als je x,y,z ook zou weten, wat zou je dan nog meer kunnen?
18. Bij welke van uw taken zou u betere ondersteuning willen, zodat deze taak gemakkelijker of interessanter wordt? Mensen optimaliseren en techniek bewaakt de grenzen.
19. Als u het voor het zeggen had, wat zou u als eerste veranderen om uw werk beter te kunnen uitvoeren?

5. Samenwerking / overig (co-operation / miscellaneous)

Zeker sinds de organisatorische splitsing van de 'oude' NS zijn er veel betrokkenen bij het proces met vaak ieder eigen ideeën en eigen belangen. Hebt u het gevoel dat u met zijn allen aan hetzelfde doel werkt? Wordt er door anderen (bv. treindienstleiders, planners, andere machinisten, bijsturing, netwerkbestuurders, …) voldoende rekening met uw werkzaamheden gehouden?

20. Wanneer, denkt u, was de klant (reiziger of vervoerder) / buurtreindienstleider / machinist blij met een door u gekozen afhandelingswijze?
21. Hoe komt u erachter of uw afhandelingswijze daadwerkelijk tot tevredenheid heeft geleid? Hoe vaak vindt dat soort terugkoppeling van klanten of collega's plaats?
22. Hoe denkt u dat treindienstleiders / netwerkbestuurders / planners *uw* werkzaamheden zien?
23. Hoe ziet u de positie van de treindienstleider / netwerkbestuurder / planner / conducteur in het proces in relatie tot uw werk?
24. Houden 'hogere' actoren voldoende rekening met uw werk? Geeft u zelf voldoende feedback? Wat gebeurt daarmee?
25. Houdt u voldoende rekening met 'lagere' actoren? Krijgt u voldoende feedback? Wat doet u daar mee?

26. Tot 1995 lag het jaarlijks aantal STS-passages ongeveer op 100 á 150 per jaar. Sindsdien is het geleidelijk aan gestegen tot zo'n 250 á 300 nu. Wat zijn volgens u mogelijke verklaringen daarvoor?
27. Nogmaals: Geef voorbeelden van situaties waarin iets niet loopt zoals u had gewild. Hoe komt dat? Wat had je vooraf moeten weten om een betere oplossing te kunnen kiezen? (beter qua veiligheid, tijdigheid van instellen, treinpunctualiteit, energieverbruik, reizigersgemak, anders).

6. Samenvattend (summary)

Kunt u, na bovenstaande, de voor u *belangrijkste* zaken nog eens benadrukken? Wat kan op korte termijn, wat op langere termijn verbeterd worden?

The same questionnaire also served as a basis for interviews with timetable designers, infrastructure designers and system architects, where inappropriate questions were skipped and some additional questions were added as appeared informative from earlier interviews. These were mainly questions of the following type:

- What was the purpose of the system under design?
- How does the design support the end user (e.g. driver, traffic controller?)
- Was the design realised as intended? If not, why?
- Are there any difficulties or unexpected problems with the design?

Appendix C Subsequent meanings of the yellow signal aspect

Table C1: Changes in the meaning of the yellow signal aspect.

Issuer	Year of change	Meaning of yellow aspect	Change
NS	1949 (signalling system 1946)	When passing the signal, decrease or regulate speed such that the train can be stopped in front of the next signal that can show stop (distance at least 1000 / 250 m[19]).	-
NS	1956 (signalling system 1954)	Reduce speed to 30 km/h. Continue to brake until the train can be brought to a halt within sighting distance of a signal that displays 'stop'.	1000 m (resp. 250 m) braking distance no longer guaranteed. Instead: driver should drive on sight towards the 'stop signal'.
NS	1965	Reduce speed to 'low speed'. Continue to brake until the train can be brought to a halt within sighting distance of a signal displaying 'stop'.	Intermediate speed of 30 km/h replaced by 'low speed', which was defined as '30 km/h, unless another value is prescribed locally, and such that the braking distance is shorter than 250 m.'
NS	1967	Idem	Low speed was now defined as '40 km/h unless indicated otherwise', thus abandoning the 250 m requirement.
NS	1975	Reduce speed to 'low speed'. Be prepared for 'stop'.	Requirement to drive on sight became implicit; emphasis is on the intermediate 'low speed'.
NS	1990	Reduce speed to 40 km/h. Depending on visibility and braking conditions, continue with such a speed – but not higher than 40 km/h – so that it is possible to stop in front of the next 'stop' signal.	Tries to make the behaviour when running at 40 km/h explicit
RailNed	1995	Limit speed to 40 km/h or so much less as to be able to stop in front of the next signal that shows 'stop'.	Requirements for driving slower than 40 km/h made result-controlled; reason for doing so is unknown.
Ministry of Transport	2004	Limit speed to 40 km/h and continue, depending on visibility and braking conditions, with such a speed that it is possible to stop in front of the next signal that orders to stop.	Meaning became 'juridified' and the criteria for running slower than 40 km/h were made explicit (yet these criteria are incomplete).

[19] Depending on high or low speed.

Appendix D Learning from SPAD databases

Calculation of SPAD probability multiplication factors using the SPAD workgroup database
The first half of this appendix will explain how the calculations for section 7.3.6 were made and why they yield only a fraction of the desired results.

Official statistics as for example collected and analysed by the railway inspectorate are not fully useful for such an analysis: although information on a high number of possible contributing factors are collected[20], these are only some of the factors as we need to fill Table 7.3 and Table 7.4 (see page 123). Therefore, for the calculation made in this thesis, a database was used from the sector-wide SPAD workgroup (section 2.2.7). This database was set up according to the method described by the SPAD workgroup (Wright and Vorderegger 2005) in order to define a list of the thousand most risky signals in the network, where a patch to the ATB system (ATB-VV, see section 9.4.3) was planned to be installed. Since it requires probabilities to calculate risks, it should be possible to use this database also for our calculation.

Yet, there are many limitations that obstruct a detailed calculation. The first and most severe limitation is that the database does not contain information on how often signals are approached with a red aspect, a problem that was also found by an ergonomic consultancy firm that was asked to analyse why freight trains pass red signals relatively often (Van der Weide et al. 2009). The resulting problem is that if, for example, the multiplication factor indicates that dwarf signals are passed at red 4 times as often as normal signals on a post, one cannot say whether this is because dwarf signals are 4 times as difficult to approach or because dwarf signals are displaying red aspects 4 times as often. Of course, any combination of these two possibilities can be the case as well. There is no other option than to assume that the percentage of red approaches is the same for all signals. For some signal types this assumption is more valid than for others. For example, it is unlikely that signals placed in a curve display red aspects significantly more often than those along a straight track section. On the other hand, dwarf signals are often used on side tracks and may therefore have a higher percentage of approaches at red (which might also lead to a lower SPAD probability if this means that the driver expects a stop). Thus, the effect of this assumption on the results is not straightforward. In order to avoid having to make this assumption, it would be necessary to measure how often signals are approached when displaying a red aspect; however, this was not possible during this study.

Secondly, the database does not contain data for all criteria mentioned by drivers, so only some of the driver's perceptions could be assessed. No data were available for a number of important criteria and their influence on the SPAD probability could not be determined. These factors were:
- sighting distance
- signal aspect sequences (e.g. yellow-yellow-red)
- number of signals related to the number of tracks
- percentage of signal approaches while the signal was showing the red aspect
- whether the distant signal coincides with a fixed speed reduction

[20] See last part of this appendix for more information about the inspectorate's analysis of SPAD incidents.

Thirdly, only some information in the database could be taken into consideration. This is not a limitation of the database itself, but necessary because not all entries in the database contained enough data to calculate the probabilities.

- From all approximately 9.800 signals in the Netherlands, it includes 6.213 signals in station areas and at junctions. The remainder are automatic block signals on the open track, which are not included in the analysis.
- Signals which are equipped with the ATB-NG train protection system are also excluded from analysis, so that 5.715 signals remain.
- The signal type was unknown or irrelevant (e.g. only for shunting) for 776 of the remaining 5.715 signals; these have been excluded from the analysis.
- Train frequencies are known for only 45,5% of all signals. Signals on yards and those on tracks normally run over in the opposite direction, were assigned a 0 trains per hour frequency. Inclusion of such signals would lead to a division by zero when calculating SPAD probabilities per approach, so these signals were left out.
- Therefore, in total, 2601 signals (about 26,5% of all signals) had sufficient data and were taken into account.

Finally, with regard to the properties of the signals that one would like to calculate the SPAD probabilities for, some fields were inadequately filled and these factors could not be taken into account in the calculations.

- There is a field that indicates how many signals are placed next to each other. This field, however, was indicated with 0 (instead of 1 or 'unknown') if there was only one signal and/or if the number was not known. This made it difficult to interpret these data, so that this attribute was not taken into account in the analysis.
- For signal placement one of the following cases could be chosen: (normal, in/behind curve, before bridge, behind bridge, before tunnel, in tunnel, behind tunnel, at platform). These categories are overlapping, thus creating a structural problem in the database, making it unsure how to handle for example a platform signal in a curve before a bridge. Apparently, these criteria were also not clear for those who filled in the database, because many possible categories were not used or at least not in all cases in which one would expect this. Therefore, the only categories taken into account in the calculation were 'normal' and 'in/behind curve', which appeared to be reliably filled in.

The information in the SPAD workgroup database is therefore not optimal, but it was the best source that was available which relates (some) signal properties, train frequencies and SPAD data. Table D1 provides an overview of the data used. The fact that gantry signals are overrepresented and dwarf signals are underrepresented in the data set can possibly be explained by the fact that dwarf signals are often used on side tracks which have no regular traffic, and thus no train frequency data. Gantry signals are almost always placed on spots with multiple tracks, which tend to be the more busy parts of the network. Indeed, usage data are available for only 35% of dwarf signals, compared to approximately 60% for gantry signals and signals on a post.

Given the relatively high unreliability of the source data, no statistical significance tests have been performed. The aim of a significance test is to check whether differences in outcomes can be explained by chance, but this assumes reliable data on the basis of which the calculation is undertaken. Since the database was incompletely and unreliably filled, the source data must be considered unreliable, so that a significance test does not make sense: even if such a test would show a statistically significant difference, one would not know

Table D1: Information on data used.

signal type	all signals (excl. ATB-NG)		signals with frequency data		representation
	number	share	number	share	in dataset (%)
signal on post	2212	38.7%	1383	53.2%	62.5%
gantry signal	515	9.0%	309	11.9%	60.0%
dwarf signal	2212	38.7%	759	29.2%	34.3%
other / unknown	776	13.6%	150	5.8%	-
total	5715	100.0%	2601	100.0%	n/a

whether this shows a significant difference in the SPAD probability or in the reliability with which the database has been filled. Thus, the data provided in this section should be taken as an indication only.

This means that the results should be interpreted with care. Further cautions in interpreting the data are:

- The calculated error probability may still have a considerable variance (that may or may not be explained by some of the factors marked with a question mark in Table 7.3 or Table 7.4), so on certain spots the SPAD probability may be far higher or lower.
- The data should mainly be considered as a relative measure, e.g. to approximate that dwarf signals on average have a 4 times higher probability of being SPADed than a signal on a normal post. In this way, multiplication factors can be derived to estimate the increase in SPAD probability for a number of signal properties.

Table D2 shows the resulting SPAD probabilities for each signal type and layout. Note that these results include also approaches at other than red aspects, so that the real SPAD probability must still be divided by the share of red aspects shown in the signal. For example, if two in a million approaches result in a SPAD, and the signal shows red for 10% of the approaches, the SPAD probability per approach at red is 2 in 100,000.

Table D2: SPAD probabilities ($\times 10^{-6}$) per approach at either one signal aspect for different signal types and layouts.

	no. of signals	average	in curve ?		<250 m behind points?	
			no	yes	no	yes
signal on post	1383	1.8	1.7	2.7	1.8	3.0
gantry signal	309	3.4	2.8	5.6	3.1	8.9
dwarf signal	759	7.9	7.4	12.0	6.9	9.9
total	2451	2.9	2.7	4.3	2.5	7.3

Learning from SPAD incidents and accidents by the inspectorate's database

One main objective of SPAD investigations is to learn what their causes are and so be able to prevent future SPADs. One of the organisations which do so is the railway inspectorate (Inspectie Verkeer en Waterstaat 2007g). An overview of all conclusions in the inspectorate's incident reports was sent to the Dutch parliament in May 2009. What is striking is that many conclusions are drawn again and again, for instance that there is insufficient guidance for rail safety during timetabling and during real time control. The fact that these conclusions return a number of times raises the question as to whether those who are addressed are sufficiently eager to improve matters, and whether the inspectorate has sufficient power to really enforce that the lessons learned are picked up, so that recurrence of similar accidents can be prevented.

Table D3: Hierarchical causes of SPAD incidents as used by the inspectorate.

Cause:	Description:
Ground procedure	Activities of ground staff violate procedures, e.g. providing a written order incorrectly
On-board procedure	Activities of train staff violate procedures, e.g. when the guard incorrectly allows the driver to depart
Technical failure	E.g. brake failure, slippery track
Traffic controller operations	The traffic controller has revoked the route
Miscommunication	A SPAD results from unclear messages exchanged between ground and train staff.
Expectancy[21]	The driver did not expect the red signal
Distraction	The driver was distracted (attention was decreased). E.g. because of a technical failure when the train approaches a red signal.
Perception of previous signal[22]	The driver has problems with the visual detection of the preceding (yellow) signal. E.g. because of poor weather conditions, the driver did not see the signal and was not prepared to stop.
Perception	The driver has problems with the visual detection of the red signal. E.g. because the signal is placed in a curve.
Brake application by driver	The signal is passed at danger because the driver had problems to stop the train (or keep it stopped). E.g. the driver brakes too late or not hard enough.

Not all SPADs lead to an incident report. For most SPADs, namely those without severe consequences, a number of entries in a standard form are filled out, so as to collect data on SPAD incidents. This approach is similar to the epidemiologic approach which was described in section 3.3.3. Then, the alleged cause of the SPAD is mapped onto a standard incident with a predefined hierarchical order of causes, which is shown in Table D3 (Inspectie Verkeer en Waterstaat 2007f). Normally, a SPAD cause will be overruled if another cause higher in the list also applies. This approach leads to attribution of accidents and incidents to one cause, where in fact nearly all incidents are multicausal and should be coded as such. Also, this is forcing data into an artificial framework, instead of using these data to try to understand what is actually going on. The inspectorate is aware of this, yet says that better analysis requires better researchers (Inspectie Verkeer en Waterstaat 2007f), whom they are apparently not willing or able to appoint.

For example, if a driver was distracted and therefore did not see the signal, and therefore applied the brakes too late, ‘distraction’ would be the primary cause of the SPAD. Although this sounds logical in this example, it is questionable whether a similar hierarchy works in all cases. Also, it is questionable how precisely the categories are defined and applied. In a SPAD accident near Weesp, a driver was distracted by the smell of a brake failure (hence the hierarchical table would indicate that the brake failure was the cause of the SPAD) but nonetheless it was concluded that the driver had a (wrong) expectancy (Inspectie Verkeer en Waterstaat 2008f).

[21] The inspectorate indicates that the subjective interpretation of the driver (or any other person concerned) of some primary causes can be decisive, e.g. ‘Expectancy’ and ‘Distraction’ cannot clearly be distinguished.
[22] This cause was added in 2005.

Thus the SPAD database is filled, but with insufficient data for proper analysis: what is missing are process-related issues. For the driver's process, it is relevant to know whether the signal is part of a yellow-yellow-red aspect sequence. Similarly, one might expect also an entry to see whether the traffic controller could have supplied a conflict-free route equally well (which was, for instance, also the case in the Weesp accident). Since such data are not included in the database, it will also never come out as a possible explanation for SPADs: any research which involves the database will necessarily mirror the 'causes' of the imposed hierarchical order, instead of providing conclusions that are actually relevant to know. This is not unique though: a Swedish study of investigation manuals of eight big companies revealed that 'what you look for is what you find' (Lundberg et al. 2009). A similar criticism was also made by Dabekaussen, who re-analysed 69 SPAD dossiers that were available at IVW. He concluded that the hierarchical framework 'will exert a strong influence on the collection and interpretation of data' (Dabekaussen 2005). Other outcomes of his research were:

- Only active errors, such as driver errors, are recorded in the inspectorate's analyses Latent factors are not considered.
- Direct causes and root causes are confounded 'making it impossible to use the analysis model for anything but the most rudimentary causal analysis, let alone the ability to use the results for determining effective countermeasures'.
- The focus is on what happened, not why that happened.
- Analyses are consequence-driven, so that SPADs with severe consequences receive a more thorough analysis than those with no consequence this time.

Dabekaussen proposed to use a systems approach (which is in line with the findings of this thesis), and to take the presence of unique or unexpected factors as a selection criterion for further analysis rather than the severity of consequences. Also in the sector-wide safety advisory board (OVS), it was mentioned several times that the incident investigation reports by the inspectorate only address a selection of possible contributing factors to an incident. The inspectorate answered that they lack time to analyse everything, so that they take some aspects into account in one report, and others in other reports. However, this selection is never made explicit in the report itself.

Although the incident analyses do lead to useful conclusions, many more useful conclusions would be within reach if causes of SPADs were to be linked to the driver's process and if coding of more than one cause was made possible and, indeed, required. Also, it seems to be necessary to look not only at whether operations comply with the rules, but also whether the rules themselves make sense. The inspectorate argued that this is not their formal task, which is correct according to their formal task definition. However, this leaves open the question who else should question the rules?

Appendix E Costs and benefits

Reluctant to quantify costs and benefits
When discussing the main argument of this thesis with ProRail employees, many of them indicated that they expected management to ask for a quantification of costs and benefits of any proposals I would make. Although this is not part of this scientific work, several other sources provide an indication of benefits that can likely be achieved.

Note that several items that should be solved in order for the railways to perform better, as they were mentioned in Chapter 9, can be difficult to quantify. What are the benefits if your company has improved knowledge of operational processes, or if staff is more motivated and thinks in a wider scope? The fact that these items are difficult to quantify *does not mean* that these are unimportant requirements to improve! This section will only provide a rough estimation of the following items:
- Energy saving
- Reduction in delay minutes
- SPAD reduction
- Process times per train movement

Energy savings
NS spends an annual amount of € 90 million for traction energy (1,5 TWh or $1{,}5\times10^9$ kWh), so the costs are € 0,06 / kWh. Each percent saved on energy consumption pays back as approximately 1 million euro (Sierts et al. 2007). The relatively big masses and high speeds of rail vehicles mean that much energy is involved in accelerating trains. Similarly, much energy can be saved when the need for re-acceleration of trains can be prevented.

A simple calculation shows that (re)accelerating an eight coach double deck intercity train (400t) to 140 km/h costs about 300 MJ (83 kWh); approximately 10% of the energy is lost during transport in the overhead line system so that 91 kWh must be supplied; the costs involved are then 5,48 euro. Note that the energy required to accelerate a train grows quadratically with the train's speed. A train that was slowed down to 100 km/h and needs to reaccelerate to 140 km/h still needs half the energy that would be necessary to accelerate it from 0 to 140 km/h. Similarly, it is often thought that heavy freight trains consume lots of energy; however, accelerating a more or less modal freight train (1200 t) to 80 km/h also takes 300 MJ. Re-acceleration of trains should therefore be avoided unless recuperative braking is widely available. If the train is diesel-powered, it takes approximately 22 litres (assuming a calorific value of 42.7 MJ/kg, density of 0.85 kg/lit and engine efficiency of 37%). Depending on the price level of Diesel fuel, costs involved are approximately 17 to 20 euro. To get all these numbers in the right perspective: it also takes also about 300 MJ to light a 9W low-energy bulb for a whole year.

Several sources provide estimates of energy savings for different solutions. Their results are summarised in Table E1. An interesting project that was not mentioned earlier in this thesis is the Energiesparende Fahrweise (ESF) which was integrated by Deutsche Bahn into the electronic timetable displays (EbuLa). It shows the driver how much energy he has consumed on a trip. Only showing this information has already led to a decrease in energy consumption; this effect was stronger when a competition was started to achieve the lowest possible energy consumption without affecting the timetable. Although somewhat out of the scope of this thesis, it is a nice example of how measuring real operations data helps to motivate staff and achieve the company's goals.

Table E1: Energy savings for several traffic management solutions.

Solution	Source	Solution Type	Saving	Benefits
Co-Production	Lüthi (2008)	rescheduling with driver assistance	10-30%	10-30 M€/yr
'de Groene Golf'	Hemelrijk (2004)	rescheduling with driver assistance	10%	10 M€/yr
ESF EbuLa	Schaal (2007)	no rescheduling, energy consumption shown to driver + competition	6%	6 M€/yr
RouteLint	Albrecht et al. (2007)	no rescheduling, only driver assistance	3-5%	3-5 M€/yr

Fewer delay minutes

The amount of train delays prevented is an interesting performance measure for different traffic management solutions. However, it does not primarily relate to the service requirements as for instance defined by the service intention. In some cases, the benefit is expressed as an increase in punctuality, which makes different calculations difficult to compare because this is dependent on the starting value of the punctuality; e.g. an increase in punctuality from 80% to 82% is totally different than an increase from 97% to 99%, although both are seen as a 2 percent point increase. In the first case, it means that 2 out of 20 previously unpunctual trains are made to run on time, or 'healed' (10%), whereas these are 2 out of 3 previously unpunctual trains in the second case (67%). Therefore, punctuality improvements have been recalculated as 'healing power'. Table E2 provides an oversight of possible improvements.

Table E2: Effect of several traffic management solutions on delay minutes and punctuality.

Solution	Source	Solution Type	'Punctuality healing power'	Decrease delay minutes	Benefits
Co-Production	Lüthi (2008)	rescheduling with driver assistance	n/a	53-78%	unknown
Co-Production	Lüthi (2008)	rescheduling without advisory speeds	n/a	10-48%	unknown
'de Groene Golf'	Hemelrijk (2004)	rescheduling with driver assistance	21%	60%	unknown
Conflictsignalering /Beslissingsondersteuning	Projectteam Definitiestudie VPT/BB (2002)	rescheduling with driver assistance	15 – 37.5%	n/a	unknown
Conflictsignalering /Beslissingsondersteuning	Makkinga et al. (2002)	rescheduling without driver assistance	17.5%	n/a	46 M€
RouteLint	Albrecht et al. (2007)	no rescheduling, only driver assistance	11%	n/a	unknown

Although these numbers come from different studies, the main picture is consistent: if no rescheduling takes place, but the driver is informed about the traffic situation, the options for improving the situation are smallest and so are the benefits (e.g. RouteLint). If rescheduling takes place, but the driver is not informed about the new plan, there are more options for improving the plan and the benefits (e.g. CS/BO without advisory speeds). The three examples with rescheduling and driver assistance show the highest improvements in the same ranges.

SPAD reduction
Reliable estimates about SPAD reduction are difficult to achieve. As was shown in this thesis, the SPAD probability is influenced by many factors, such as the characteristics and education of the driver, the timetable design, the fluidity of the traffic process, the information exchange between traffic controller and driver, the signalling systems and its layout. Measuring (near)accident data during the research would not produce enough information to be statistically reliable and creation of artificial data, as could be done by means of simulation, would require many man-years of simulation, since every driver is involved in a SPAD every 10 to 20 years on average, which would be an even longer time span if solutions are very effective. However, a prerequisite for a SPAD is to approach a signal at red. From automated simulation studies it was found that fictitious drivers who have a coasting driving method run into approximately 1,5 (unplanned) signals at red per simulation run, compared to 2,5 for drivers who are running at full speed all the way (Kohls and Watson 2008). This alone leads to a 40% reduction. Also, the Swiss Co-Production approach reverses the problem as it tries to make sure that trains will not arrive at a yellow distant signal before it has cleared to green. It must be expected that the amount of (unexpected) approaches at red will then be close to zero. A 50% reduction in red signal approaches seems then even a conservative guess. This will therefore also lead to approximately a 50% reduction in SPAD rates (currently approximately 200 yearly cases in the driving process; see section 1.2.3), and given a 1:100 ratio between SPADs and SPAD accidents (see Figure 1.1 and Figure 1.2), this approach prevents approximately 100 yearly SPADs and 1 SPAD accident.

Few publicly available sources exist that provide information on costs associated with SPADs. For some SPAD accidents, estimates are sometimes available in the public investigation reports. These normally only include direct costs such as repair costs for rolling stock and infrastructure, but no costs for accident investigation, replacement bus services, image loss or treatment of injuries. As can be seen from Table , the amount of damage easily adds up to 2 million euro per accident, so if one annual SPAD accident can be prevented, savings will be at least 2 million euro per year. Introducing a signal aspect for ‘short braking distance’ was calculated to cost ƒ 7.2 million (price level 1987). Assuming a 4% inflation rate and with the conversion rate for euro’s (2,20 Dutch guilders = 1 euro) the change would now cost 8 M€, or the equivalent of 2 to 4 prevented accidents.

Table E3: Costs associated in SPAD accidents.

Accident location	Source	Rolling stock damage	Infrastructure damage	Benefits
Roosendaal	Inspectie Verkeer en Waterstaat (2005a)	4 M€	Unknown	> 4.0 M€
Arnhem	Inspectie Verkeer en Waterstaat (2007c)	1.4 M€	0.2 M€	> 1.6 M€
Amsterdam Muiderpoort	Inspectie Verkeer en Waterstaat (2007b)	1.2 M€	1 M€	> 2.2 M€

Process time reduction
With regard to a higher network utilisation, the key is to decrease process times so that the production elements become available for a new task after a shorter time. As was shown, the current way of coping with disruptions by providing slack time is inefficient because at the 'moment suprême' the operators do not know exactly how much slack is available. Note that slack is usually a minute or so. This appears to be only a small amount of time, but this is often a deception and this one minute represents very valuable time:

- The blocking time for a train is approximately two minutes (see Table 2.2); with current practices, a one minute buffer time between paths is usual. It appears to be possible to gain several paths per hour if the required amount of buffer time can be reduced. Only a few sources quantify how much the required amount of time per train movement decreases with more precise operations. Swiss tests have shown that it is possible to operate a train within a 15 second time window, so that 45 seconds per train could be gained compared to the Dutch buffer of approximately 60 seconds. This means that for every three trains that are scheduled now (3×180 sec + running time differences = 540 sec + running time differences), a fourth one can be scheduled (4×135 sec + running time difference = 540 sec + running time differences). This requires that planning and operations are being planned at a level of seconds, so that a driver can see whether he is approaching the boundary of his 15 second window.
- If there is a one minute running time supplement in a train movement between two succeeding stations, this can easily add up to five or more minutes for the whole trip and to ten minutes for a round trip from A to B and back. In the case of a ten minute frequency, as is striven for with the current future plans for the Dutch network, this means that the ability to remove slack saves one train set (worth typically approximately 2 – 12 million Euros depending on train length and train properties) and its staff. Also, it creates a faster service and thus possibly a higher income (Hofstra 2005).

With track access charges of approximately €0,7 per trainkm + €1,774 per 1000 tkm (ProRail 2009b), the value of an additional train path is approximately €1,00 - €8,00 per km. However, it is difficult to estimate the exact amount of benefits since additional train paths also involve additional costs for administration and track wear, and the precise income depends on the type of train that will use the additional paths and the distance it will travel.

Costs and benefits
Table E4 summarises the savings that were calculated above. Costs for a traffic management system and driver support system were calculated in 2002 with regard to the BB21 project (Projectteam Definitiestudie VPT/BB 2002) and amount in the most advanced version and including risks in development to 15 M€ for a traffic management module and an additional 2 M€ for a driver support system (in the latter it is not clear whether this also involves on-board equipment).

Note that further improvements are likely to be achieved through further reduction of process times or through signalling improvements. Some other measures, such as staff motivation, appear to be available almost for free. Also analysis of performance data will probably pay off: this is also common practice in a commercial sector such as aviation; also the German energy saving example shows that it works. Knowledge of what is actually going on is a first prerequisite to improve. However, the costs or benefits of these measures are difficult to quantify; partly because their effect is difficult to quantify and partly because their effect depends on other measures already taken: for example, if better traffic management already leads to 50% fewer red signal approaches, the benefits of improved signalling will also decrease. These effects have therefore been left out of Table E4. Despite this, it is clear that

the development of a traffic management system pays off within 0.5 to 2 years from energy benefits alone. All associated safety and capacity benefits then come to all intents and purposes for free. It would be interesting to measure these in practice so as to increase knowledge about the real effects of traffic management.

Table E4: Summary of costs and benefits.

Costs		Benefits		
Conflict detection and resolution	< 15 M€	Energy saving	10-30%	10-30 M€/yr
Driver assistance (time/advisory speed)	< 2 M€	Delay minutes	-60%	unknown
		SPAD reduction	-50%	>2 M€ /yr
		Process time saving	-25%	unknown
Total	< 17 M€			>> 12-32 M€/yr

Glossary

ARI	Automatic Route Setting system (Automatische rijweginstelling)
ATB	Automatic Train Protection system (Automatische treinbeïnvloeding)
ATB-NG	New Generation ATB
AV	Design rule (Algemeen Voorschrift)
DDA	Detect – Diagnose – Act
distant signal	Signal that indicates a preview of its corresponding main signal at braking distance.
ETCS	European Train Control System. European standardised interface for the exchange of train contol information between track, train and driver.
EU	European Union
FHT	Average repair time (Functiehersteltijd)
interlocking	Device to administer infrastructure use and lock routes whilst they are in use.
KPI	Key Performance Indicator
local traffic controller	Traffic controller who sets route in one station or on one line.
main signal	Signal that gives access to a track section.
NMa	Netherlands Competition Authority (Nederlandse Mededingingsautoriteit)
NS	Nederlandse Spoorwegen (biggest public rail transport company in the Netherlands)
NS'54	Current Dutch signalling system.
Plan(ning)	A plan is a number of coherent measures in order to achieve certain goals. Planning is a (methodical) tool for control. Planning can be described as taking a coherent series of decisions such that the course of events will be influenced (de Leeuw 2001).
PRL	Local traffic management module of the VPT system (Procesleiding).
Process control	The process of causing a system variable to conform to some desired value, called a reference value (Franklin et al. 2006).
Quality	The degree to which a set of inherent characteristics fulfils requirements (ISO9000 2000).
Real time	A process is controlled in real time if the reaction time in response to an event is shorter than the maximum time imposed by the process that must be controlled. Real time control is necessary to avoid that the control system lags behind the process it controls. It therefore serves to avoid instability (Encyclopædia Britannica Online 01 Sep. 2009).
Real time process control	In this thesis mainly used to indicate the work of the traffic control centres and train drivers.

regional traffic controller	Traffic controller who reschedules traffic and closes line sections in case of emergency.
Reliability	The ability of an item to perform a required function, under given environmental and operational conditions and for a stated period of time (ISO8402 1986).
Route	Physically reserved infrastructure sections for a specific train (thus being the operational equivalent of a train path). Note that a *path* has dimensions in time and space; a *route* has only a spatial dimension since it can only exist as of the moment that it has been set.
Safety	Freedom from those conditions that can cause death, injury, occupational illness, or damage to or loss of equipment or property (MIL-STD-882D 2000).
Service	Non-material equivalent of physical goods. A product or a service is a (temporary or permanent) change in the customer and/or the context, which fulfils a need. A service is intangible, cannot be stored, often there is coproduction (the customer coproduces, otherwise no production is possible). A service can only be assessed after it has been delivered and the controllability of the result is only limited (de Leeuw 2001).
signal	Communication device to send messages to the driver (or other staff).
signal aspect	Message that is being communicated by a signal.
SPAD	Signal Passed At Danger
TAD	Document that prescribes how to deal with delayed trains (Treinafhandelingsdocument).
TAO	Train service affecting irregularities (treindienstaantastende onregelmatigheid)
TOC	Train Operating Company
Tolerance	The allowable deviation from the desired reference value.
TRAIL	Research School for Transport, Infrastructure and Logistics
Train path	A planned trajectory in time and space in which infrastructure is reserved for a specific train movement. A path is a service delivered by an Infrastructure Manager to a Train Operating Company.
TVP	Maintenance slot on the infrastructure (treinvrije periode)
VKL	Regional traffic management module of the VPT system (Verkeersleiding).
VPT	Planning and control software for timetable design and traffic management (vervoer per Trein).

Summary

Modelling Risk Control Measures in Railways
Analysing how designers and operators organise safe rail traffic
Jaap van den Top

Research question and relevance
The Dutch government and the railway industry wish to use the existing railway network more intensively. This will likely involve more interactions between trains, which could result in delays or in trains incidentally passing red signals more frequently than is currently the case. It is the main focus of this thesis to develop knowledge that helps to prevent these phenomena from happening.

Better understanding of how to cope with these interactions is required to see how the governmental wish can be realised, whilst the quality and safety levels remain acceptable. If the current performance level is not to be lowered and if disturbances cannot be taken away, the only way to achieve a higher network utilisation is to improve the control system's ability to cope with disturbances. Since these interactions are inherent to operational processes and how these are controlled, the research question is:

In what way could real time process control be improved to positively influence safety and quality of service in rail traffic operation?

Method used
To answer the research question, the following main steps were taken:
1. basic concepts from railway operation and safety science were analysed and gaps identified;
2. a framework, the so-called 'safe envelope of operations', was developed which can incorporate the main issues relevant to railway operation and safety science;
3. interviews and literature surveys were taken to identify to which extents current practices match the criteria from the framework;
4. options for improvement were explored.

Analysis of basic concepts from railway operation indicated that many functions have to co-operate in order to supply train paths. These are mainly infrastructure design, timetable design, traffic management, route allocation and train control. The cascade model can show these as a hierarchical set of functions. It indicates that in the current situation, degrees of freedom for operators become less as the time remaining to the moment of operation gets shorter. Also, it indicated that there is hardly any feedback from the lower-level functions towards the design stages.

Literature from safety science indicated that risks and benefits are two sides of the same activity: they cannot be considered in isolation, but must be seen in the context of normal operations. Also, it was shown that most existing accident models do not suffice to describe certain types of accidents in relation to normal system operation; they focus on safety only. Also, many accident models are based on probabilities of component failure or operator failure. Although useful, such models do not explain which mechanisms underlie systemic risks and what must be done to decrease the given probabilities. This requires understanding of the underlying processes.

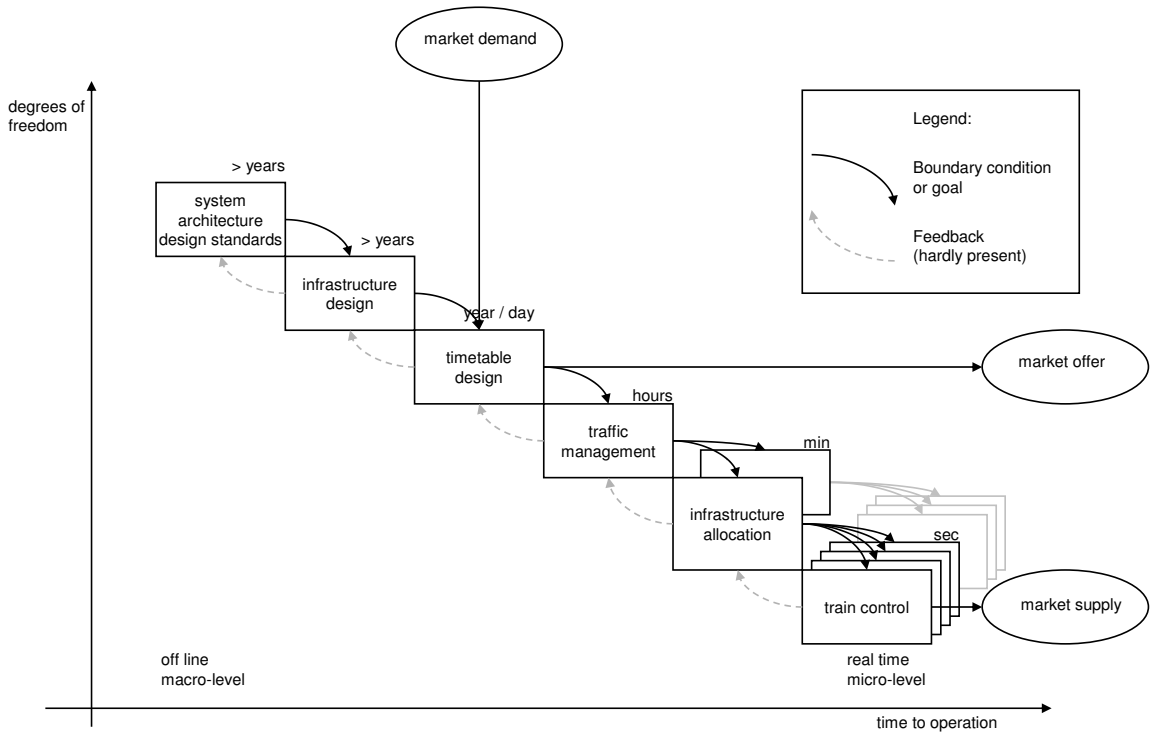


Figure S1: The cascade model.

Therefore, in this thesis, several existing and new concepts were combined and developed in order to understand these processes. This so-called 'safe envelope of operations' is a risk model that can combine safety risks with the purpose of an activity. It can represent operational processes and its critical boundaries and as such can help designers, managers and analysts to understand how risks and quality problems arise. The model indicates that actors controlling a system should be able to know the following items in order to fulfil their tasks safely and effectively:

- What is the current system state?
- What is the goal and which tolerances are allowed (operational envelope)?
- Can the goal be reached in time (attainability envelope)?
- Where is the boundary of safe operation (viable envelope)?
- What is the ultimate moment to exercise control (controllability envelope)?

These five elements can be visualised relative to each other, and each of these items can change over time: the envelopes are dynamic. Thus, the safe envelope model provides additional ways to think of possible accident mechanisms in dynamic systems. To keep the system within the safe envelope's boundaries, control must be exercised by the operators. Modelling this requires understanding of the technology, the humans operating the system, but also of the designers and managers who are responsible for the system as a whole. Although it appears that quite different fields of knowledge are involved, it was shown that concepts from the different fields can all be related to control loops:

- Control theory from the field of engineering;
- Detect-diagnose-act control sequence from safety science;
- Situation awareness from the field of human factors;
- Plan-do-check-adjust feedback loop from management science;
- Organisational learning from the field of management science.

It is convenient that the same principles can be applied to these different fields. This allows us to describe both safety and 'production' by taking an integrated functional approach to the system's variables that need to be controlled. This requires a clear visualisation of the boundaries within which the variables should be (this is the function of the safe envelope model) and it requires a model to describe the control processes that must keep the variables within these boundaries (for which the detect-diagnose-act sequences were used). All functions in the control model (as for example described in the cascade model) need to detect where the envelope's boundaries are, to diagnose the situation and to perform an action to influence the system state. Such interference in the system can be done either directly, or indirectly when the action provides a changed goal, a changed boundary, a new control measure to a lower level control functions, or if these are given a buffer or barrier to absorb disturbances.

In order to assess the controllability of the current railway system, several sources were used, comprising interviews with operational staff and designers, and literature from scientific and business sources. Since problems will show themselves in operations, this was the starting point for the analysis. Wherever necessary, the study looked back to earlier choices, for example in the timetable design or in systems design. To find underlying causes in the detailed data, a grounded theory approach was used.

Results
Twelve related underlying causes were found for performance problems (e.g. delays or passing red signals), which are shown in Figure S2. These underlying causes can be grouped as three vicious circles of causes and effects:

A. *Individual loop*
   Since no clear goal is set (1), it is not really clear for many operators what exactly they should achieve. This either results in self-formulated goals (3) and self-developed practices (10), or to operators limiting themselves to their formal responsibilities (2). For several reasons, many operators are discouraged (7) and thus provide no feedback (8). Without feedback, there is insufficient organisational learning (9). Problems remain which further discourages operators.

B. *Operations loop*
   Also, this causes operators to develop individual and therefore possibly incompatible methods (10), which are partly a cause of inadequate control information and the loss of motivation. This starts a self-reinforcing process of operators retreating to their own 'islands' with limited communication between them (12).

C. *Design and management loop*
   This also makes it difficult for designers to understand how they can support operators with better support systems, regulations et cetera (11). This again leads to a shift from goal to means (3) and support systems that do not always adequately provide the operators with the information they need (4,5,6)

There is very little feedback on the basis of which actors in the cascade model can learn whether their decisions work in practice. This may be one of the reasons why there are different representations of the reality for different functions. For instance, in timetable design, traffic control and train driving, five different and in some situations incompatible ways of solving train delays were found. These five ways are all based on certain assumptions to make the task of the actor in question easier (at least for himself), but they do not always fit

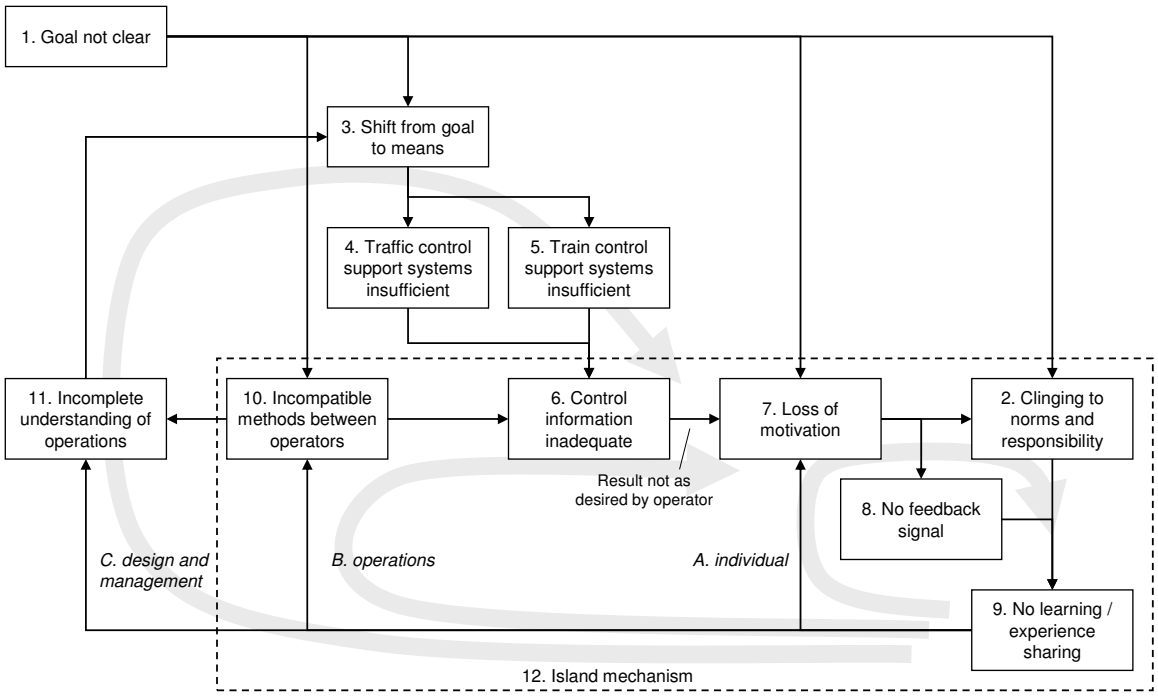


Figure S2: Cause-effect diagram for underlying causes of performance problems.

together. Many operators retreat to their own island in order to make their task manageable, but by retreating, they may cut off others from relevant information.

With regard to the traffic control process it is remarkable that this process is difficult to predict, even though the traffic controller's main function is to solve conflicts in advance.
It would be helpful if a traffic controller had tools that provide him with an oversight of the situation as it is expected to be in 15 – 60 minutes and if conflicts in that picture can easily be found. The effort of a traffic controller would then not be spent on finding out what the situation actually is and where problems are about to occur, but on selecting solutions to solve these problems in advance. Although operator experience partially fills the gap of incomplete information, the result it that normally no updated, conflict-free, real time schedule is made in response to smaller deviations from the timetable. Thus, it is not guaranteed that there is a feasible production schedule, a fact that itself makes the process unpredictable.

Also the drivers have difficulties to perform their main task, which is to operate their train closely to their scheduled path. Apart from the fact that new conflict-free schedules are normally not made in response to smaller disruptions, the driver is also not informed about possibly changed schedules. Also, he cannot see in sufficient detail whether he does adhere and how he can adhere better to the (re)scheduled path, which requires a precision in fractions of minutes. These facts result in an unnecessarily high number of restrictive (yellow or red) signal approaches. Thus, preventing that signals are passed at danger is not just a human factor or technical (signalling) issue, but also an information exchange issue between traffic controllers reworking the plan and drivers executing it so that trains can be controlled such that they meet signals preferably only when they have cleared the route behind them.

Although timetable designers and traffic controllers usually rely on the driver to obey signals, imperfections in the route setting system, the signalling system and the train protection system, and the impossibility of faultless operator behaviour make this a dangerous assumption. Driver expectancy cannot be banned, since this is essential human behaviour. The focus should not be on prevention of expectancies, but effort should instead be spent to

preventing that these expectancies can become wrong, which is also an information exchange issue. One of the main problems is that the signalling system in several cases provides information that does not inform the driver about the unique scenario that he is going to enter, which leaves room for (wrong) interpretation. This puts stronger emphasis on the driver's ability to detect a red signal from a sufficient distance and relate it to his track, but it was shown that this is not sufficiently well possible in some cases.

Route knowledge, which is in fact also an expectancy, is in these cases the last line of defence in order to assure safety, but is only has a weak formal meaning. Moreover, the human memory works more reliably in response to external cues and should thus not be used to replace these. Therefore, although route knowledge is useful as general experience, it should not be relied on to ensure safe operations.

The belief that driver expectancy is the cause of many problems gives the impression that the system design is failure-free, and does not need to be changed, but this is not the case. Several examples were shown in which changes to the traffic control system, the signalling system, the train protection system or the radio system made the driver's task more difficult in an unexpected way. Since this was not the designers' intention, it must be concluded that their influence on the driver's task is not fully understood. Often, new technology was introduced without a clear view on how this would precisely support the operator's processes, or new technology replaced older technology and thereby, informal working routines were rendered infeasible.

Possible improvements
In the first place, the wish to 'better' use the system requires a clear cut answer to the question when the use is 'good', or in other words, what the system should achieve. The second question that should be asked is how the operational processes must be designed in order to achieve this result. This asks for redesigned processes in the first place, such as improved communication between traffic controllers and train drivers and improved situation awareness. This again will often require new technology in order to support the processes. The intended operating process should be designed first and the required technology should be derived from that, not vice versa.

The production processes, their planning and the means of production, such as the timetable, the rolling stock capabilities, the infrastructure layout and the safety system limitations must be geared to each other in order to fully exploit their possibilities. This asks for an integrated approach: if one of these elements is considered as a given and immutable fact, improvements in the other elements cannot be fully exploited. Since process control defines when which resources will be used for which process, it is inherently linked to the achievement of the goals and thus to the system's capacity.The current paradigm in timetable design, in which it is assumed that trains do not interact and a unique solution to operate them is nailed down a year in advance, does not take the effect of process control into account and therefore needs to be replaced. What is required instead, is a new timetabling paradigm in which the traffic pattern is structured in such a way that the real time process control has more and explicit freedom to reschedule the traffic pattern on the short term. The customer service intention, which should remain explicit, can then serve as a clear goal for the traffic control process, so that the quality of service is likely to be improved. Better operator support would steer the system more efficiently. This would then allow for a more efficient use of slack time and the ability to cut on slack times, thus leading to a higher utilisation. Since the system would then be operated closer to the boundary of what it can handle, constant monitoring of whether operational processes are running as desired and without side-effects would be necessary.

This means that not only real time traffic control is necessary, but also the traffic control and train control processes themselves must be monitored and analysed.

On an abstract level, solving quality and safety problems is all about avoiding surprises (undesired, unintended and unexpected results) during operations. To improve (safety) performance, the correct functioning of the processes performed needs to be made sure. Such insights can be gained by closing the feedback loop from operational level to design and management. Problems during operations should be freely reported by operators and actively be looked for by the organisation. For example, the organisation must measure how often safety-critical processes are executed, how often they failed and how often failures were stopped by the safety systems. Examples are the number of route setting commands refused by the interlocking, or the number of interventions by the train protection system. Such indicators make clear how well the operational process is actually kept away from its critical boundaries. Such a process-related analysis allows the finding of more possible causes for trains running through red signals than an approach in which causes are considered to be certain properties of signals or of drivers, since it also takes their interaction into account. As long as no proper analysis of current operations takes place, knowledge about the actual processes will not improve and neither will the processes themselves. Feedback of actual performance data to timetable design has recently been successfully started with; it is recommended that such feedback loops are also closed on other disciplines.

In spite of the errors addressed earlier, the current signalling system, in combination with a high level of professionalism of train drivers and traffic controllers, has resulted in railways being one of the safest transportation systems. However, if the public demands on the system increase and trains have to be planned even more closely together, the system will be taxed beyond its limits. Right now there are insistent signs, in the form of operators losing oversight over the situation and increasing numbers of signals passed at danger, that indicate that the limits of the current control philosophy are approaching. Therefore, a much more sophisticated real-time control philosophy is required, which must enhance operators' situations awareness so that all actors share the same expectation. The signalling system can then, again, be used as a backup system – the function it was designed for. Failures and unexpected surprises should be looked for and freely reported, so that operational processes can be geared to each other and so that designers can better understand how actual operations are being performed. This will not only improve safety performance, but it will also make it more likely that the promises to the customer can be fulfilled.

Samenvatting

Risicobeheersing van spoorwegen modelleren
Onderkennen van de risico's van railverkeer in ontwerp en praktijk
Jaap van den Top

Onderzoeksvraag en relevantie
De Nederlandse overheid en de spoorsector beogen het bestaande spoorwegnet intensiever te benutten. Waarschijnlijk heeft dit tot gevolg dat treinen elkaar ten opzichte van de huidige situatie al bij kleinere afwijkingen van de dienstregeling in de weg zullen zitten, hetgeen kan leiden tot vertragingen of (incidenteel) tot het onterecht passeren van een stoptonend sein. Het ontwikkelen van kennis die ertoe bijdraagt om deze fenomenen te voorkomen vormt de focus van dit proefschrift.

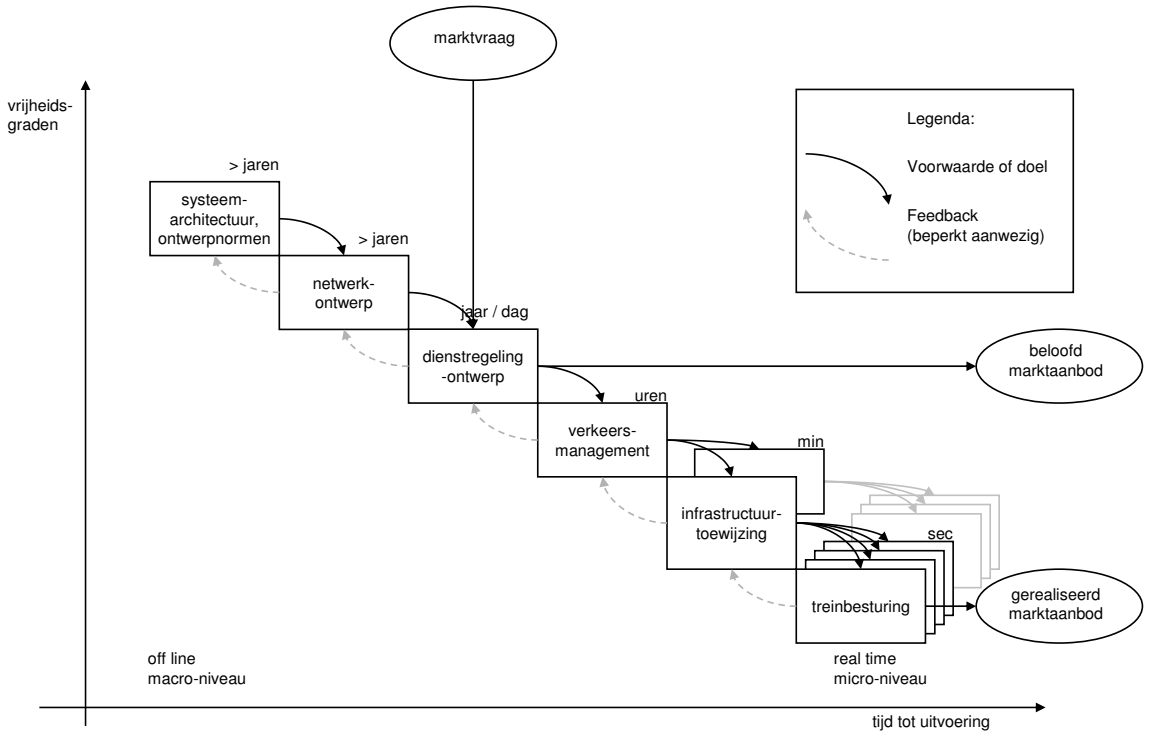
Om de wens van de overheid en de spoorsector te kunnen realiseren zonder dat het kwaliteits- of veiligheidsniveau achteruitgaat, is een beter inzicht nodig in hoe met deze relatief kleine verstoringen kan worden omgegaan. Wanneer het kwaliteits- en veiligheidsniveau niet mogen verslechteren, en wanneer de verstoringen niet kunnen worden weggenomen, is de enige wijze om een betere benutting te realiseren door beter met deze verstoringen om te gaan. Deze verstoringen vinden per definitie plaats in de uitvoering, zodat de onderzoeksvraag luidt:

Op welke wijze kan de procesbesturing in de uitvoering worden verbeterd zodat het veiligheids- en kwaliteitsniveau positief wordt beïnvloed?

Gevolgde werkwijze
Om de onderzoeksvraag te beantwoorden zijn op hoofdlijnen de volgende stappen gezet:
1. basisbegrippen van spoorwegexploitatie en veiligheidskunde zijn bestudeerd en gaten in kennis geïdentificeerd;
2. een theoretisch kader, de zogeheten 'veilige begrenzing voor de procesuitvoering' is ontwikkeld, waarin de relevante zaken voor spoorwegexploitatie en de veiligheidskunde tezamen kunnen worden weergegeven;
3. interviews en literatuuronderzoek hebben plaatsgevonden om vast te stellen in hoeverre de huidige werkwijze voldoet aan de eisen die het theoretisch kader stelt;
4. verbetermogelijkheden zijn verkend.

Om treinpaden – reserveringen op het spoorwegnet voor een treinbeweging – te kunnen leveren, moeten verschillende functies samenwerken, zoals infrastructuurontwerp, dienstregelingontwerp, verkeersleiding, rijweginstelling en treinbesturing. Deze functies kunnen in samenhang worden weergegeven in het in dit proefschrift geïntroduceerde cascademodel (afbeelding S1). Hieruit blijkt dat, naar mate het moment van uitvoering dichterbij komt, er steeds minder vrijheid overblijft voor hen die het systeem besturen. Het model geeft tevens aan dat er nauwelijks sprake is van feedback van de operationele functies naar de ontwerpfuncties.

Veiligheidskundige literatuur geeft aan dat veiligheidsrisico's en de beoogde voordelen van een activiteit twee zijdes van een medaille zijn: deze kunnen niet los van elkaar worden beschouwd. Het bleek echter ook dat veel veiligheidskundige modellen niet voldoen om bepaalde ongevalstypen in relatie tot de normale procesuitvoering te beschrijven; zij focussen uitsluitend op veiligheid, of ze zijn gebaseerd op de faalkansen van een systeemcomponent of



Afbeelding S1: Het cascademodel.

van menselijk handelen. Hoewel dat zinvolle informatie kan opleveren, geven zulke modellen niet aan welke mechanismen ten grondslag liggen aan het ontstaan van risico's en wat moet worden gedaan om de als gegeven beschouwde kansen te verlagen. Dit vereist begrip van de onderliggende processen.

Met dit doel zijn in dit proefschrift verschillende bestaande en nieuwe veiligheidsmodellen gecombineerd en doorontwikkeld. Het zogeheten model van de 'veilige begrenzing voor de procesuitvoering' (safe envelope of operations) is een veiligheidsmodel dat veiligheidsrisico's en de beoogde doelen van een activiteit kan combineren. Het visualiseert de operationele processen en de kritieke grenzen waarbinnen deze moeten blijven, en kan zo ontwerpers, managers en analisten inzicht geven hoe risico's en kwaliteitsproblemen ontstaan.

Het model geeft aan dat de actoren die een systeem besturen de volgende informatie moeten hebben om hun taken effectief en veilig te kunnen vervullen:
- Wat is de huidige toestand?
- Wat is het doel, en welke toleranties daarop zijn toegestaan? (operationele begrenzing)
- Kan dit doel tijdig worden bereikt? (bereikbaarheidsbegrenzing)
- Wat is de grens voor schadevrij opereren? (levensvatbaarheidsbegrenzing)
- Wat is het laatst mogelijke moment waarop nog een effectieve besturingsmaatregel kan worden genomen? (bestuurbaarheidsgrens)

Het model kan deze vijf elementen en hun onderlinge samenhang visualiseren. Elk van de elementen kan gedurende de tijd veranderen; ze zijn dus dynamisch. Daarmee geeft het model aanvullende manieren om na te denken over mogelijke ongevalsmechanismen in veranderende systemen. Om het systeem binnen de gestelde begrenzingen te houden, moeten de bestuurders ervan maatregelen uitoefenen. Om dit te modelleren is het nodig niet alleen te begrijpen hoe de techniek en de mens in het systeem werken, maar ook te kijken naar de rol van de ontwerpers en managers die verantwoordelijk zijn voor het functioneren van het systeem als geheel. Hoewel dit geheel verschillende vakgebieden lijken, blijkt dat concepten

uit elk van de vakgebieden op een of andere manier als regelkring kunnen worden beschreven, hetgeen handig is voor verdere analyse. Het betreft:
- Meet- en regelkringen uit de ingenieurswetenschappen;
- Waarnemen – diagnosticeren – handelen uit de veiligheidskunde;
- Situatiebewustzijn uit de cognitieve psychologie;
- Deming's plan-do-check-adjust regelkring uit de bedrijfskunde;
- Organisatorisch leren uit de bedrijfskunde.

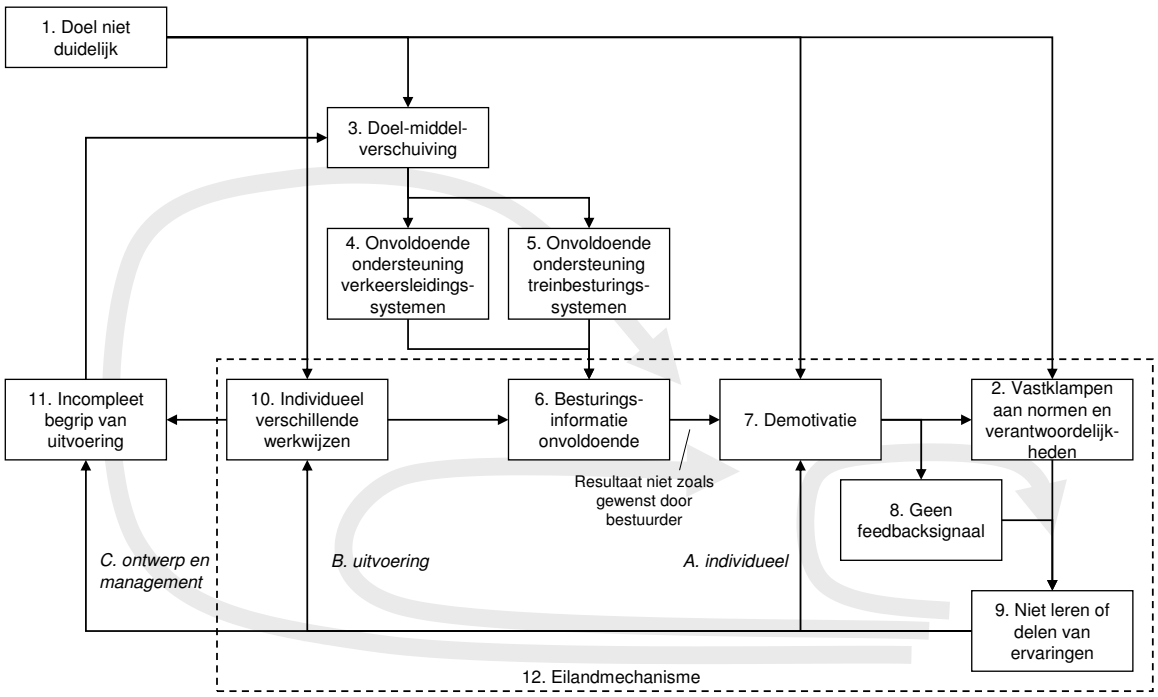
Veiligheid en 'productie' kunnen dan beide worden beschreven met een integrale, functionele beschrijving van de systeemvariabelen die moeten worden beheerst. Dit vereist een heldere visualisatie van de grenzen waarbinnen de variabelen zich moeten bevinden (dit is de functie van het model van de veilige begrenzing). Daarbij is tevens een model nodig dat de besturingsprocessen kan beschrijven, waarvoor de opeenvolgingen van waarneming – diagnose – handeling kunnen worden gebruikt. Alle functies in het besturingssysteem (zoals bijvoorbeeld beschreven in het cascademodel) moeten waarnemen waar de begrenzingen zijn, de situatie beoordelen en een actie uitvoeren om de systeemtoestand te beïnvloeden. Dat kan rechtstreeks, of via tussenliggende besturingsfuncties door deze te voorzien van een nieuw doel, een nieuwe begrenzing, door deze functies een nieuwe besturingsmaatregel te verschaffen, of deze van een buffer of barrière te voorzien die verstoringen kan opvangen.

Om de bestuurbaarheid van het huidige spoorwegsysteem te beoordelen, zijn verschillende bronnen gebruikt. Bijvoorbeeld wetenschappelijke artikelen en vakliteratuur, maar ook interviews met ontwerpers en vooral uitvoerend personeel. Omdat problemen zich uiteindelijk voordoen in de uitvoering, was deze het startpunt van de analyse. Waar nodig werd in deze studie teruggekeken naar eerdere (ontwerp)keuzes, zoals naar het ontwerp van de dienstregeling en naar het ontwerp van technische systemen. Om uit deze aldus verkregen gedetailleerde informatie achterliggende oorzaken te vinden werd de methode van *grounded theory* gevolgd.

Resultaten
Er werden twaalf onderliggende oorzaken gevonden voor de prestatieproblemen, zoals vertragingen of het passeren van een stoptonend sein. Deze onderliggende oorzaken kunnen worden gegroepeerd in drie vicieuze cirkels van oorzaken en gevolgen, zoals is weergegeven in afbeelding S2:

A. *Individuele cirkel*
   Omdat er geen duidelijk doel is gesteld (1) is het voor het uitvoerend personeel niet altijd duidelijk wat er precies bereikt moeten worden. Dit leidt ofwel tot door henzelf geformuleerde doelen (3) en individueel verschillende werkwijzen (10) of zij klampen zich vast aan normen, regels en verantwoordelijkheden (2). Om verschillende redenen zijn veel uitvoerenden gedemotiveerd (7) en zij geven daarom geen feedback of verbetertips (8). Zonder dergelijke feedback kan de organisatie echter niet leren en verbeteren. De problemen blijven in stand, wat weer een bron is voor demotivatie.

B. *Uitvoeringscirkel*
   Tevens leidt dit ertoe dat het uitvoerend personeel individueel verschillende, en daarom mogelijkerwijs onsamenhangende, werkwijzen ontwikkelt (10). Dit is een van de oorzaken van onvoldoende informatie over het te besturen proces (6) en ook daaruit vloeit demotivatie voort. Dit leidt tot een zichzelf versterkend proces waarin het uitvoerend personeel zich terugtrekt op een overzichtelijk 'eilandje' met een beperkte mate van onderlinge communicatie (12).



Afbeelding S2: Oorzaak-gevolg-diagram voor onderliggende oorzaken van prestatieproblemen.

C. *Ontwerp- en managementcirkel*
De impliciete en heterogene werkwijze in de uitvoering maakt het moeilijk voor ontwerpers om het proces precies te begrijpen (11) en te beoordelen hoe zij het proces beter kunnen ondersteunen. Dit leidt vaak tot een doel-middel-verschuiving en de ondersteunende systemen bieden niet altijd de informatie die het uitvoerend personeel nodig heeft (4,5,6).

Er is inderdaad zeer weinig terugkoppeling op basis waarvan actoren in het cascademodel kunnen leren of hun beslissingen in de praktijk ook werken zoals beoogd. Dit kan een van de redenen zijn waarom er verschillende representaties van het operationele proces zijn: zo zijn er bijvoorbeeld tijdens het dienstregelingontwerp, de verkeersleiding en de treinbesturing vijf verschillende manieren gevonden om rekening te houden met treinvertragingen. Deze zijn stuk voor stuk gebaseerd op bepaalde aannamen, die weliswaar het werk van de actor in kwestie eenvoudiger maken (althans voor hemzelf) maar die niet altijd op elkaar passen. De actoren trekken zich terug op een eiland om hun taak overzichtelijker te maken, maar door dit te doen, snijden ze anderen af van relevante informatie.

Aan het verkeersleidingsproces valt op dat het voor betrokkenen moeilijk is om een toekomstverwachting van dit proces te maken; dit ondanks het feit dat het de taak van de verkeersleiding is om conflicten juist op voorhand te voorkomen. Het zou beter zijn wanneer verkeersleiders een beeld zouden hebben van de verwachte verkeersafwikkeling in de komende 15 – 60 minuten en dat conflicten in dat beeld zouden worden aangegeven. Een verkeersleider zou dan zijn energie kunnen steken in het bedenken van een oplossing voor een probleem, in plaats van te moeten zoeken of, waar en wanneer zich probleem een probleem voordoet. Hoewel ervaring en vakmanschap een deel van dit gat opvullen, wordt er door de verkeersleiding normaliter geen up-to-date, conflictvrij plan gemaakt als er kleine afwijkingen op de dienstregeling optreden. Het is daardoor niet zeker dat er een maakbare ‘productieplanning’ is, een feit dat op zichzelf weer eraan bijdraagt dat het proces zich onvoorspelbaar gedraagt.

Ook de machinist ondervindt problemen bij het uitvoeren van zijn hoofdtaak, die eruit bestaat om de timing van de trein netjes te laten overeenstemmen met het geplande tijdpad. Nog los van het bovengenoemde feit dat verkeersleiding niet steeds het plan uitvoerbaar en up to date houdt, wordt de machinist doorgaans ook niet op de hoogte gesteld van een vernieuwd plan. Evenmin beschikt hij over middelen die hem in voldoende detail kunnen weergeven of hij netjes volgens dit pad rijdt; dit vereist een precisie op fracties van minuten. Als gevolg van deze oorzaken komen treinen onnodig vaak seinen tegen met restrictieve seinbeelden (geel of rood). Voorkomen dat treinen door rood rijden is daarom niet alleen een vraagstuk met betrekking tot menselijk falen of technische vangnetten, maar ook een vraagstuk over informatie-uitwisseling tussen de verkeersleiding en machinisten, opdat treinen zodanig kunnen worden bestuurd dat zij pas bij seinen aankomen wanneer deze doorrijden toestaan.

Hoewel dienstregelingontwerpers en verkeersleiders erop vertrouwen dat machinisten hun seinbeelden opvolgen, is dit een gevaarlijke aanname vanwege onvolkomenheden in het systeem voor automatische rijweginstelling, het seinstelsel, het systeem voor automatische treinbeïnvloeding, en ook omdat menselijk handelen nooit foutvrij kan zijn. Een verwachtingspatroon kan niet worden afgeschaft; het is een essentieel onderdeel van menselijk functioneren. De focus zou daarom niet moeten liggen op de wens om een verwachtingspatroon op zichzelf te voorkomen, maar op het voorkomen dat de verwachting onjuist wordt. Ook dit is een informatie-uitwisselingsvraagstuk. Een probleem daarbij is dat het seinstelsel in een aantal situaties niet op unieke wijze kan aanduiden in welk scenario de machinist terecht zal komen, hetgeen ruimte biedt aan (verkeerde) verwachtingen. Dit legt meer druk op de juiste waarneming van een rood seinbeeld en de mogelijkheid om dit aan de te berijden rijweg te relateren, maar het bleek dat ook dit in bepaalde situaties niet goed mogelijk is.

In die gevallen is wegbekendheid, in feite ook een verwachtingspatroon, het laatste middel om de veiligheid te waarborgen, maar wegbekendheid heeft nauwelijks een formeel geregelde inhoud. Bovendien werkt het menselijk geheugen effectiever in reactie op een externe prikkel; wegbekendheid zou dus niet moeten worden gebruikt om deze prikkel te vervangen. Wegbekendheid als ervaring is weliswaar nuttig, maar zou niet noodzakelijk moeten zijn om een veilige procesuitvoering te waarborgen.

Het geloof dat een verwachtingspatroon de oorzaak is van veel problemen, geeft de indruk dat het systeemontwerp foutloos is en niet veranderd hoeft te worden, maar dat is niet het geval. In dit proefschrift zijn verschillende voorbeelden gegeven waarin veranderingen in het verkeersleidingssysteem, het seinstelsel, het treinbeïnvloedingssysteem of het radiocommunicatiesysteem de taak van de machinist op onverwachte wijze hebben bemoeilijkt. Aangezien dit niet de bedoeling van de ontwerpers is geweest, moet worden geconcludeerd dat hun invloed op het operationele proces door hen niet volledig wordt begrepen. Vaak werd nieuwe techniek ingevoerd zonder dat voldoende duidelijk was op welke wijze deze het proces moest ondersteunen, of verving nieuwe techniek oudere techniek en maakte daarbij informeel ontstane werkwijzen onmogelijk.

Mogelijke verbeteringen

De wens om te komen tot een ‘betere’ benutting van de spoorwegen roept allereerst de vraag op wat dan onder een ‘goede’ benutting wordt verstaan. Met andere woorden: wat moet er eigenlijk worden bereikt? De tweede vraag die zou moeten worden gesteld, is op welke wijze de operationele processen moeten worden ingericht zodat dit resultaat kan worden bereikt. Om deze nieuwe processen beter te kunnen ondersteunen, kan het nodig zijn om nieuwe

technische mogelijkheden aan te wenden. Het is dus nodig eerst het proces te ontwerpen en de benodigde techniek daarvan af te leiden.

De productieprocessen, en dus ook de planning daarvan en de productiemiddelen waarmee deze worden uitgevoerd, moeten op elkaar worden afgestemd, om hun mogelijkheden ten volste te kunnen benutten. Als voorbeeld kunnen dienen de dienstregeling, de mogelijkheden van het rollend materieel, de infrastructuurlayout, en de beperkingen vanuit de beveiligingsinstallaties. Dit vereist een geïntegreerde aanpak: als één van deze elementen als een onveranderbaar gegeven wordt beschouwd, kunnen verbeteringen in één van de andere elementen niet altijd worden benut.

Aangezien de procesbesturing bepaalt wanneer welke middelen voor welk proces zullen worden gebruikt, is de wijze van procesbesturing op inherente wijze verbonden aan het al dan niet bereiken van de gestelde doelen en dus van invloed op de capaciteit van het systeem. Het huidige paradigma in het dienstregelingontwerp, waarin wordt aangenomen dat treinen geen onderlinge interactie vertonen en een unieke oplossing voor de verkeersafwikkeling dus een jaar van te voren kan worden vastgesteld, moet daarom worden losgelaten. In plaats daarvan moet een nieuw paradigma komen, waarin het dienstregelingontwerp de treinenloop structureert, en wel zodanig dat de belofte aan de klant kan worden nagekomen (externe functie), hetgeen tevens een helder doel schept voor het verkeersleidingsproces (interne functie). Daarbij kan het verkeersleidingsproces meer en expliciete ruimte worden gegeven om nog op korte termijn te beslissen over de best mogelijke treinvolgorde. Als het uitvoerend personeel beter ondersteund wordt, kan het systeem efficiënter worden bestuurd. Daardoor kunnen marges en buffers in de dienstregeling zinvoller worden gebruikt en gedeeltelijk worden verminderd, zodat een hogere benutting mogelijk wordt. Het systeem ‘draait’ dan wel dicht bij de grenzen van het haalbare, zodat het nodig is om continu te monitoren of het systeem doet wat het moet doen, en zonder onbedoelde bijwerkingen. Dit betekent dat niet alleen het verkeersleidingsproces het real-time proces moet monitoren, maar dat ook het juist functioneren van de verkeersleidings- en treinbesturingsprocessen zelf moeten worden gemonitord en geanalyseerd.

Om de (veiligheids)prestatie te verbeteren, moet dus de juiste afloop van de activiteiten worden verzekerd. Inzicht daaromtrent kan worden verkregen door terugkoppeling van operationeel niveau naar het niveau van ontwerp en management. Zij kunnen dan vaststellen of alles werkt zoals bedoeld. Op een meer abstract niveau is het oplossen van kwaliteits- en veiligheidsproblemen vooral een kwestie van het voorkomen van (ongewenste, onbedoelde en onverwachte) verrassingen tijdens de uitvoering. Problemen tijdens de uitvoering zouden vrijelijk moeten worden gemeld door uitvoerend personeel, terwijl de organisatie er ook actief naar moet zoeken. Zo zou de organisatie bijvoorbeeld moeten bijhouden hoe vaak een veiligheidskritisch proces wordt uitgevoerd, hoe vaak een dergelijk proces faalt en hoe vaak dan het vangnet in werking treedt. Concrete voorbeelden zijn het aantal keren dat de beveiligingsinstallatie verhindert dat een rijweg voor een trein wordt ingesteld, of het aantal ingrepen door het automatische treinbeïnvloedingssysteem. In beide gevallen moet dit worden gerelateerd aan het aantal keren dat dit proces wel succesvol, dat wil zeggen zonder ingreep, verloopt.

Dergelijke indicatoren maken inzichtelijk hoe succesvol de procesbesturing erin slaagt om het systeem binnen de veiligheidsgrenzen te laten werken. Een dergelijke procesgerelateerde analyse leidt ertoe dat er meer mogelijke oorzaken voor bijvoorbeeld het passeren van een stoptonend sein worden gevonden; meer dan bij een aanpak waarin oorzaken van dergelijke problemen worden beschouwd als eigenschappen van ofwel de seinen ofwel de machinisten.

Deze spelen weliswaar ieder een rol in het proces, maar de verwevenheid tussen beide moet niet uit het oog worden verloren. Zo lang er geen structurele analyse plaatsvindt van de feitelijke procesgang, kan de kennis erover niet verbeteren en het proces zelf dus ook niet. Met de terugkoppeling van realisatie naar dienstregelingontwerp is inmiddels een succesvol begin gemaakt, dat ook naar andere terreinen zou kunnen worden uitgebreid.

Ondanks de problemen die in dit proefschrift zijn aangegeven, slaagt het huidige seinstelsel in combinatie met het vakmanschap van verkeersleiders en machinisten erin de spoorwegen een van de veiligste vervoerwijzen te maken. Wanneer echter de maatschappelijke eisen aan het systeem, en daarmee ook de benutting, steeds meer worden opgevoerd, kan het systeem overbelast raken. Er zijn nu al signalen in de vorm van toenemende aantallen stoptonendsein-passages en personeel dat het overzicht verliest. Dit duidt erop dat de grenzen van het huidige besturingsparadigma nabij zijn. Voor een hogere benutting is daarom bovenal een goed doordachte procesregeling nodig, die het situatiebewustzijn van het uitvoerend personeel zodanig verbetert dat alle actoren dezelfde toekomstverwachting delen. Het seinstelsel kan dan weer, zoals het oorspronkelijk bedoeld is, worden gebruikt als vangnet voor de gevallen dat er nog een onvolkomenheid in het plan zit. Fouten en onverwachte verrassingen moeten worden gezocht en gerapporteerd, zodat de processen op elkaar kunnen worden afgestemd en ontwerpers meer inzicht kunnen krijgen in de wijze waarop het proces daadwerkelijk verloopt. Niet alleen kunnen zulke inzichten de veiligheid verbeteren, maar ook kunnen zij ertoe leiden dat de beloften aan de klant beter kunnen worden nagekomen.

Curriculum Vitae

Jaap van den Top was born in Arnhem on November 18, 1980. He obtained his gymnasium diploma from Stedelijk Gymnasium Arnhem in 1999. Given his broad and deep interest in railway operation, he studied civil engineering at Delft University of Technology and graduated cum laude in 2005. Topic of his M.Sc. thesis was to analyse the possibilities of dynamic rail traffic management in the station of Arnhem, which he did during an internship at ProRail.

When Jaap was about to finish his M. Sc. Thesis, ProRail and TU Delft again joined forces in the Next Generation Infrastructures research programme. Since the new project fitted well onto his prior M.Sc. thesis, Jaap took up a PhD position at TU Delft's Safety Science Group.

Currently, Jaap is working as an accident investigator, one of his activities being the investigation of railway accidents.

NGInfra PhD Thesis Series on Infrastructures

1. Strategic behavior and regulatory styles in the Netherlands energy industry
   Martijn Kuit, 2002, Delft University of Technology, the Netherlands.
2. Securing the public interest in electricity generation markets, The myths of the invisible hand and the copper plate
   Laurens de Vries, 2004, Delft University of Technology, the Netherlands.
3. Quality of service routing in the internet: theory, complexity and algorithms
   Fernando Kuipers, 2004, Delft University of Technology, the Netherlands.
4. The role of power exchanges for the creation of a single European electricity market: market design and market regulation
   François Boisseleau, 2004, Delft University of Technology, the Netherlands, and University of Paris IX Dauphine, France.
5. The ecology of metals
   Ewoud Verhoef, 2004, Delft University of Technology, the Netherlands.
6. MEDUSA, Survivable information security in critical infrastructures
   Semir Daskapan, 2005,Delft University of Technology, the Netherlands.
7. Transport infrastructure slot allocation
   Kaspar Koolstra, 2005, Delft University of Technology, the Netherlands.
8. Understanding open source communities: an organizational perspective
   Ruben van Wendel de Joode, 2005, Delft University of Technology, the Netherlands.
9. Regulating beyond price, integrated price-quality regulation for electricity distribution networks
   Viren Ajodhia, 2006, Delft University of Technology, the Netherlands.
10. NetworkedrReliability, Institutional fragmentation and the reliability of service provision in critical infrastructures
    Mark de Bruijne, 2006, Delft University of Technology, the Netherlands.
11. Regional regulation as a new form of telecom sector governance: the interactions with technological socio-economic systems and market performance
    Andrew Barendse, 2006, Delft University of Technology, the Netherlands.
12. The Internet bubble - the impact on the development path of the telecommunications sector
    Wolter Lemstra, 2006, Delft University of Technology, the Netherlands.
13. Multi-agent model predictive control with applications to power networks
    Rudy Negenborn, 2007, Delft University of Technology, the Netherlands.
14. Dynamic bi-level optimal toll design approach for dynamic traffic networks
    Dusica Joksimovic, 2007, Delft University of Technology, the Netherlands.
15. Intertwining uncertainty analysis and decision-making about drinking water infrastructure
    Machtelt Meijer, 2007, Delft University of Technology, the Netherlands.
16. The new EU approach to sector regulation in the network infrastructure industries
    Richard Cawley, 2007, Delft University of Technology, the Netherlands.
17. A functional legal design for reliable electricity supply, How technology affects law
    Hamilcar Knops, 2008, Delft University of Technology, the Netherlands and Leiden University, the Netherlands.
18. Improving real-rime train dispatching: models, algorithms and applications
    Andrea D'Ariano, 2008, Delft University of Technology, the Netherlands.

19. Exploratory modeling and analysis: A promising method to deal with deep uncertainty
Datu Buyung Agusdinata, 2008, Delft University of Technology, the Netherlands.
20. Characterization of complex networks: application to robustness analysis
Almerima Jamaković, 2008, Delft University of Technology, Delft, the Netherlands.
21. Shedding light on the black hole, The roll-out of broadband access networks by private operators
Marieke Fijnvandraat, 2008, Delft University of Technology, Delft, the Netherlands.
22. On stackelberg and inverse stackelberg games & their applications in the optimal toll design problem, the energy markets liberalization problem, and in the theory of incentives
Kateřina Staňková, 2009, Delft University of Technology, Delft, the Netherlands.
23. On the conceptual design of large-scale process & energy infrastructure systems: integrating flexibility,reliability, availability,maintainability and economics (FRAME) performance metrics
Austine Ajah, 2009, Delft University of Technology, Delft, the Netherlands.
24. Comprehensive models for security analysis of critical infrastructure as complex systems
Fei Xue, 2009, Politecnico di Torino, Torino, Italy.
25. Towards a single European electricity market, A structured approach for regulatory mode decision-making
Hanneke de Jong, 2009, Delft University of Technology, the Netherlands.
26. Co-evolutionary process for modeling large scale socio-technical systems evolution
Igor Nikolić, 2009, Delft University of Technology, the Netherlands.
27. Regulation in splendid isolation: A framework to promote effective and efficient performance of the electricity industry in small isolated monopoly systems
Steven Martina, 2009, Delft University of Technology, the Netherlands.
28. Reliability-based dynamic network design with stochastic networks
Hao Li, 2009, Delft University of Technology, the Netherlands.
29. Competing public values
Bauke Steenhuisen, 2009, Delft University of Technology, the Netherlands.
30. Innovative contracting practices in the road sector: cross-national lessons in dealing with opportunistic behaviour
Mónica Altamirano, 2009, Delft University of Technology, the Netherlands.
31. Reliability in urban public transport network assessment and design
Shahram Tahmasseby, 2009, Delft University of Technology, the Netherlands.
32. Capturing socio-technical systems with agent-based modelling
Koen van Dam, 2009, Delft University of Technology, the Netherlands.
33. Road incidents and network dynamics, Effects on driving behaviour and traffic congestion
Victor Knoop, 2009, Delft University of Technology, the Netherlands.
34. Governing mobile service innovation in co-evolving value networks
Mark de Reuver, 2009, Delft University of Technology, the Netherlands.
35. Modelling risk control measures in railways
Jaap van den Top, 2010, Delft University of Technology, the Netherlands.

Order information: info@nginfra.nl