top margin: 15mm

Jeroen de Ridder

author: Thesis Sans Regular 30pt, 70% Pantone Black 7

# Reconstructing Design, Explaining Artifacts

title: Thesis Sans Black 30pt, Pantone 382

## Philosophical Reflections on the Design and Explanation of Technical Artifacts

right margin: 15mm

left margin: 15mm

subtitle: Thesis Sans Regular 30pt, 70% Pantone Black 7 (btw, line space is 36pt for everything)

banner 80% Pantone Black 7, starts at 50% of page height

*Simon Stevin Series in the Philosophy of Technology*

series title: Thesis Sans Italic 12pt, Pantone Black 7, 10 mm from top of banner

banner ends at 75% of page height

series title goes here should the banner contain an image (not the case with this volume)

bottom margin: 20mm

RECONSTRUCTING DESIGN, EXPLAINING ARTIFACTS

PROEFSCHRIFT ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft
op gezag van de Rector Magnificus prof.dr.ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op vrijdag 15 juni 2007 om 12.30 uur
door Geert Jan DE RIDDER,
bestuurskundig ingenieur,
geboren te Leerdam

# **Reconstructing Design, Explaining Artifacts** *Philosophical Reflections on the Design and Explanation of Technical Artifacts* by Jeroen de Ridder

Delft 2007

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. P.A. Kroes
Prof. dr. ir. A.W.M. Meijers

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter
Prof. dr. ir. P.A. Kroes, Technische Universiteit Delft, promotor
Prof. dr. ir. A.W.M Meijers, Technische Universiteit Eindhoven en Technische Universiteit Delft, promotor
Prof. dr. P. Lipton, University of Cambridge
Prof. dr. E. Weber, Universiteit Gent
Prof. dr. Th.A.F. Kuipers, Rijksuniversiteit Groningen
Prof. dr. I. Horváth, Technische Universiteit Delft
Dr. P.E. Vermaas, Technische Universiteit Delft

Dr. P.E. Vermaas heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

*Typografie omslag en binnenwerk: Hans Stol*

Simon Stevin Series in the Philosophy of Technology
Delft University of Technology & Eindhoven University of Technology
Editors: Peter Kroes and Anthonie Meijers
Volume 4

Copyright 2007 by Jeroen de Ridder
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior permission in writing of the publisher.

ISBN 978-90-9021903-5
ISSN 1574-941X

# Table of Contents

Acknowledgements 9

**1 Introduction 13**
1.1 A Short Story 13
1.2 Two Perspectives on Artifacts: Different but Related 14
1.3 Problems and Questions 16
1.4 Background 20
1.5 Some Limitations 23
1.6 A Bit of Methodology 24
1.7 Outline and Reading Guide 26

**2 Design, Function, and Other Key Notions 29**
2.1 Introduction 29
2.2 A Philosophical Account of Design 29
2.3 Assumptions About Technical Functions 33
2.4 Another Look at Artifact Design and Explanation 36
2.5 Dispositional and Categorical Descriptions of Artifacts 42
2.6 Concluding Remarks 46
Appendix: A Worry About Circularity Dismissed 47

**3 Design Methodology: What It Does and Does not Do 51**
3.1 Introduction 51
3.2 A Brief Historical Survey of Design Methodology 52
3.3 Herbert Simon: Design as Rational Problem Solving 57
3.4 Donald Schön: Design as Reflective Practice 61
3.5 John Gero: Function, Behavior, Structure 65
3.6 Design Methodology and the Philosophy of Technology 72
3.7 Simon, Schön, and Gero Revisited 78
3.8 Conclusion 80

**4 Functional Decomposition 83**
4.1 Introduction 83
4.2 Examples of Functional Decompositions 85
4.2.1 Stamping Out and Packing Carpet Tiles 85
4.2.2 Coffee Grinder 87
4.2.3 Splashgard 88
4.2.4 Disposable Razor Blades 90
4.2.5 Lessons About Functional Decomposition 92
4.3 Further Developments in Functional Decomposition 94
4.3.1 Functional Representation 94
4.3.2 Functional Modeling 96
4.3.3 Concluding Remarks 99
4.4 Dispositional and Categorical Descriptions of Artifacts 100
4.4.1 The Dispositional-Categorical Distinction Revisited 101
4.4.2 Applying the Dispositional-Categorical Distinction 102
4.4.3 Abstractness of Functional and Structural Descriptions 108
4.4.4 When Functional Decomposition Stops 111
4.4.5 Concluding Remarks 112
4.5 Behaviors and Events 112
4.6 The Many Senses of Functional Decomposition 113
4.7 Intermezzo: Level Generation and Event Supervenience 119
4.8 More Senses of Functional Decomposition 121
4.9 What Good Is Functional Decomposition? 125

**5 Theories of Scientific Explanation 131**
5.1 Introduction 131
5.2 Preliminary Remarks and Limitations 132
5.3 Classifying Theories of Explanation 135
5.3.1 Salmon's Classification 135
5.3.2 An Improved Classification 137
5.4 Expectability 139
5.5 Pattern Subsumption 141
5.5.1 Epistemic Pattern Subsumption 141
5.5.2 Ontic Pattern Subsumption 146
5.6 Causation 149
5.6.1 Salmon's Causal account of Explanation 149
5.6.2 Other Causal Accounts 153
5.7 Relevance 158
5.8 Concluding Remarks 159

**6 Explaining the Behavior of Artifacts 161**
6.1 Introduction 161
6.2 Examples of Artifact Explanations 163
6.2.1 Paperclip 164
6.2.2 Phonograph or Speaking Machine 166
6.2.3 X-Y Position Indicator 169
6.3 Preliminary Observations and Two Explanatory Strategies 171
6.3.1 The Thing Explained 171
6.3.2 The Thing Explaining It 173
6.3.3 Two Explanatory Strategies 174
6.3.4 Features and Refinements of the Two Strategies 179
6.4 Two Strategies, not One 181
6.4.1 Preliminaries 182
6.4.2 Two Arguments for the Autonomy of Top-Down Explanations 184
6.4.3 An Additional Pragmatic Argument 187
6.5 Caveats and Objections 189
6.6 Mechanistic Explanation 192
6.6.1 Overview 192
6.6.2 Some Critical Remarks 193
6.7 Conclusion 197
Appendix: The Legitimacy of High-Level Explanation 198

**7 Intentionality and Normativity in Artifact Explanations 203**
7.1 Introduction 203
7.2 Intentionality and Action 204
7.3 Kroes on Technological Explanations 207
7.4 Kroes's Arguments Reconsidered 210
7.4.1 Two Minor Points of Criticism 210
7.4.2 The Real Trouble and a Constructive Proposal 212
7.5 A Constructive Proposal: Having vs. Explaining Functions 215
7.5.1 Relations Between the Two Projects 215
7.5.2 Artifact Explanations of Technical Functions Reconsidered 219
7.6 Delivering the Goods 221
7.6.1 A Theory of Technical Functions 221
7.6.2 Combining the Modified ICE-Theory and Artifact Explanations 224
7.6.3 Connecting Function and Structure the Right Way 226
7.7 Conclusion 227

**8 Conclusion 231**
8.1 Looking Back 231
8.2 The Dual Nature of Technical Artifacts Revisited 235
8.3 Artifact Kinds and Other Suggestions for Further Research 240

References 247

Summary in Dutch 257

Curriculum Vitae 265

# Acknowledgements

I remember well that in one of the explorative chats I had before starting the Ph.D. project of which this thesis marks the completion, my then future and now former supervisor Pieter Vermaas suggested that, since analytical philosophy of technology was such a young and unexplored field, I might well go on to become the Karl Popper, Thomas Kuhn, or Paul Feyerabend of the field. Of course, this shining prospect exercised no small appeal to an aspiring Ph.D. student.

I regret to have to admit it, but if there is one thing of which I have become quite convinced during the past five years, it is that I will not be the Karl Popper, Thomas Kuhn, or Paul Feyerabend of analytical philosophy of technology. I do not mean this in a disappointed or disillusioned sense. On the contrary, it is just that I have learned that working in a relatively young field of philosophy with little or no traditions of its own is not only exciting and full of promise, but also more difficult than I could have foreseen. With little or no default positions and counter-positions, no arguments developed to a high level of sophistication, and no general agreement on which problems are worth analyzing further, I often felt like I got thrown in at the deep end.

It was a good thing, then, that I was surrounded by people who helped and stimulated my work on this thesis in many ways. I would like to take this opportunity to thank them. First of all, my day-to-day supervisor Pieter and my official (not-quite-day-to-day-but-still-pretty-often) supervisor Peter. To accurately express their influence on this thesis I would have had to include a vast amount of footnotes throughout thanking them for ideas, suggestions for improvement, criticisms, and so on. Although this would have communicated clearly how much I owe to them, I figured it would also make for a pretty annoying read. For that reason, I express my gratitude to them here: Thank you for reading and commenting on my draft papers and chapters, stimulating me to explore new places – both intellectually and geographically –, for discussions and chats, fresh ideas to develop, incisive feedback, for friendly encouragements when I needed them, and everything else that I cannot think of right now. Your out-of-the-box thinking and independent way of doing philosophy set an example for me.

The rest of the Dual Nature group also provided an invaluable work environment, especially when it was still complete during the first couple of years of my project. Thanks Anthonie, Maarten, Wybo, and Marcel for giving me a flying start, for solving some of the hard problems before I even thought of them, and simply for creating the exceptional opportunity of being part of a group of like-minded philosophers working on strongly related problems, and everything good that came of that.

Then my fellow Ph.D. students Anke, Marcel, Lotte, Maarten, Noëmi, and Dingmar. It was good to have people around with a sense of shared destiny (which sometimes felt more like shared misfortune). Although I have not been able to make up my mind about which of the two were most enjoyable, our dinners or the discussions over it, the combination surely was excellent. Lotte deserves another mention for being such a nice person to share an office with. Thanks for listening to my expressions of optimism, pessimism, excitement, frustration, and whatnot – which surely must have been more than a little tedious at times. For stories, chats, and jokes about almost anything from religion and world-views to broken staplers and printer jams.

Thanks also to my other Delft colleagues for being a bunch of hospitable, varied, but most of all good-natured philosophers.

Next, the people who made it possible for me to study and work abroad. Michael Strevens, who gave me the opportunity to visit Stanford University during Spring Quarter 2004. Not only was his feedback on my papers highly useful and did I collect a lot of ideas about explanation of artifacts during his seminar, I also had a really good time in California. Next, Peter Lipton made my stay at the Department for History and Philosophy of Science at the University of Cambridge during Lent Term 2006 possible. While I was there, Tim Lewens read and commented on my papers and discussions about functional decomposition and artifact kinds sharpened my thinking on these topics. The Ph.D. students at Stanford and Cambridge were very kind in welcoming me and helping me find my way. Discussions with them also stimulated my thinking.

There were various reading groups, workshops, and conferences at which I had the opportunity to present and discuss my work with others. I am thankful to the audiences at them for questions and critical remarks. The regular workshops of the Dutch-Flemish Network for the Philosophy of Science and Technology warrant special mention. The old monastery of Soeterbeeck was a beautiful place for intellectual discussion and good fun. I will miss the late-night chats with Nikki, Kai, Sabina, and many others.

And then some people who do not fit in one of the above categories. Erik Weber read and offered perceptive comments on early versions of chapters 4 and 5 of this thesis. Jan Dietz shared his perspective on functional decomposition with me. My ideas about design were much enriched and corrected by

participating in two groups of industrial design students who were working on concrete design assignments. Thanks to them and their instructors.

I would also like to thank the Netherlands Organisation for Scientific Research for their generous financial support of the Dual Nature project and hence my work.

Hans Stol did a magnificent job in formatting this book and designing its cover – even though I had to reject his proposal for a catchier title for reasons of propriety – thereby relieving me of these, to my mind, herculean tasks and earning my eternal thankfulness.

On a more personal note, I would like to thank my parents and in-law family for their unflagging support to what, in their eyes, must have often been a fairly abstruse pursuit. Family, friends and members of the youth group at church provided very welcome distractions from work.

Nelleke, I know you are not into this sort of thing, but not mentioning you here would be grossly unfair. I am not going to try to express how good it is to share our lives and how it has helped my work on this thesis. Let it suffice to say that I am deeply grateful for your love and support.

Finally, thank you to our daughter Joa for enlivening the last couple of months that I was working on this thesis (not to mention a fair amount of nights during those months). I suppose this will not become the first book you read and it will certainly not be the most entertaining one, but perhaps the few pictures in it will at least capture your attention.

# Introduction

## 1.1 A Short Story

Imagine the following unlikely but instructive scenario. You are building a small garden shed. As a professional philosopher, this is not something you normally do, nor do you possess the aptitude for this sort of activity. Luckily, your friends Angela and Bob have come out of their way to help you out. At one point, Angela digs up a wooden folding ruler from her toolbox. Bob apparently has never seen such a thing before and asks her what it is. (At which point you start to wonder whether calling on Bob was such a great idea after all.) 'A folding ruler,' Angela answers matter-of-factly. Bob disagrees: 'But aren't these just some pieces of wood with regular marks on them connected by metal thingies?' 'Sure,' says Angela, slightly confused, 'but that's what a folding ruler is, right? Whoever initially came up with the idea of a folding ruler chose these materials and put them together so that the thing is now, indeed, a folding ruler. And if you don't mind I'm now going to use it to make sure these planks are all of equal length.' Things get even odder when Bob retorts: 'Surely you must be mistaken, because all I see there are just wooden sticks and pieces of metal holding them together. That's all there is and I don't see any rulers.' By now, you are seriously worried about the prospects for your shed. To lighten things up you crack a few favorite philosophical jokes. (The one about how Descartes walks into a bar, is asked if he cares for a drink, replies: 'I think not,' and, with a poof, vanishes into thin air.) Everyone agrees that philosophical jokes are pretty cheesy and with raised spirits the three of you finish the shed.

There is something distinctly weird about Bob's reactions. A natural diagnosis is to say that any disagreement he takes himself to have with Angela rests on false presuppositions. His claim about what the folding ruler is made of in no way excludes Angela's claim that the item is a wooden folding ruler. Describing an object in terms of its physical make-up is in no way at odds with describing it in terms of what it is for. Bob mistakenly assumes the opposite when he wants to push his claim to the exclusion of Angela's.

In fact, we should grant a stronger point. Not only are the two perspectives not in competition but they complement each other. Either perspective on its own leaves out something important about the ruler. A description of its physical structure leaves it indeterminate what the ruler is for, while a

description of its function and uses does not tell you much about the physical structure behind this function. This point generalizes to other technical artifacts (i.e., roughly, objects that have been designed, made, and are used for specific practical purposes). To fully understand what sort of object a folding ruler or any other technical artifact is you must take into account both its physical make-up and its function and use.

So we have a basic intuition that technical artifacts have, so to speak, two faces. For the sake of brevity and clarity, let us call the perspective with functions, purposes, goals, and use *intentional* and the other perspective in terms of structure, physical make-up, shapes, and configuration *physical*. Whether a characterization of an artifact belongs in one or the other is determined by whether or not that description contains or presupposes intentionality. With these convenient labels we can state the basic intuition as follows:

> (DN) Technical artifacts exhibit some sort of duality, consisting in their equally belonging in an intentional and a physical perspective.

This claim inspired the Delft-based research program 'The Dual Nature of Technical Artifacts,' of which this thesis is one of the last official results.[1] The following chapters will home in on the relation between the two perspectives by looking at how an artifact's having a technical function – as a paradigmatic intentional characterization – is related to its having a particular physical make-up – as a paradigmatic physical description. Before I introduce the central questions, however, I will make a couple of remarks to further flesh out the differences between the two perspectives, especially as they pertain to an artifact's technical function and its physical structure. After that, I look at how the two perspectives are related and then we will be in a position to appreciate the main questions for this thesis.

## 1.2 Two Perspectives on Artifacts: Different but Related

First, then, some points to show the differences between the two perspectives. Claims related to an artifact's technical function allow for normative and evaluative judgments in ways that claims about its physical perspective do not. To see this, compare artifacts with natural objects. 'This is a good stone' makes no sense while 'This is a good slingshot' is clearly meaningful. Or again, 'This is the worst electron I have ever seen' is nonsensical, but 'This is the worst electron microscope I have ever seen' evidently is not.[2] A special case of evaluative

[1] For more information, see Kroes and Meijers (2002; 2006) and the program's website, *www.dualnature.tudelft.nl*.

judgments are claims about malfunctioning. Artifacts can be evaluated with regard to how well or badly they perform the function they were made for, or whether they perform it at all. That is because an artifact's function sticks to it, even when it has lost the capacity to perform that function and sometimes even when it has never had the capacity for doing so. We say that a broken fountain pen still is for writing. And even though a defective car will not get you where you want to be, we readily affirm it is for driving. It is also normal to ascribe functions to malfunctioning early prototypes. None of this happens within the physical perspective. Considered exclusively as a physical object and described in terms of its physical make-up a folding ruler is for nothing in particular, just as twigs in a forest are for nothing. And save for, perhaps, colloquialisms in a physics laboratory, it is hard to imagine anyone seriously uttering 'This proton is seriously defective.'

Taking a broader view of the problem reinforces the differences between the two perspectives. The physical perspective fits smoothly into a physicalistic (materialistic) picture of the world. Seen as bare physical objects, technical artifacts are on a par with rocks, rivers, and radiation; they do not have a mental life of their own. The intentional perspective does not sit so well with a physicalistic picture of the world – that, if anything, must be a lesson of the ever-expanding literature in the philosophy of mind. The two perspectives on technical artifacts are associated with two broader and hard-to-reconcile perspectives on people: the mental and the physical.

To back up the idea that technical functions indeed belong in the intentional perspective, consider what is required for ascribing technical functions to artifacts. As a bare minimum, the existence of social practices in which material objects play a role may suffice to ascribe technical functions to the objects used. It might be argued that a flat piece of rock that is collectively used by a tribe of Neanderthals for dissecting mammoths on has the function of a table. Even if this minimalist account is right, however, it is obvious that the technical functions of most modern-day artifacts presuppose stronger forms of intentionality. The technical artifacts that furnish our life space have been purposely designed by a designer for specific purposes and are subsequently used by groups of people for those same purposes and sometimes co-opted for different purposes. Computers, cars, and cans have been intentionally designed and created to be used in specific ways to realize specific goals. To describe them as having a technical function is to tacitly assume these things. Thus, paradigm

[2] Of course anyone is free to impose some sort of instrumental context on natural objects and then evaluate them as a means to the imposed end. 'This is a good stone for breaking that window.' But that is cheating. This context must be added as a bonus, whereas artifacts are embedded in an intentional context *by definition*.

cases of technical functions presuppose intentions, goals, and actions of both designers and users. In addition, authors who have written about technical functions all agree that technical functions presuppose intentionality in some sense (Millikan 1984, 1993; Neander 1991a, b; Preston 1998b; Scheele 2005; Vermaas and Houkes 2006; Houkes and Vermaas 2006). None of this holds for the physical perspective. Physical descriptions contain no references, explicit or implicit, to intentions, goals, or actions, nor do they presuppose them in any way.

In spite of these differences, the two perspectives are also clearly related. Artifacts themselves are proof of that because in them the two perspectives are somehow brought together. This is testified by how descriptions of an artifact's function and its physical make-up mutually constrain each other. Not any arbitrary structure can realize a given function and a given structure cannot have any old function. If something has the function of a paperweight (to use an unavoidable example), it must be heavier than paper but not too heavy for people to pick up conveniently. If something is made of rubber, is tubular, and has a diameter of just over half a meter, it is a safe bet that this thing is not for, say, lighting a room or operating your DVD-player. Evaluative judgments about artifact also show that the two perspectives are related. Even though such judgments stem from the intentional perspective, as I argued above, they necessarily pertain to the artifact's physical structure as well, since it is that physical structure that performs well or badly.

What is far from clear, however, is how exactly the two perspectives are related and, specifically, how an artifact's having a technical function and its having a particular physical structure are related. Notions like intention, purpose, function, artifact behavior, and structure all have a role to play in our understanding of technical artifacts, but it is hard to see how they all fit together. There are no straightforward inferences from descriptions of technical functions to physical descriptions or the other way around. Nor is it possible to simply translate one type of description into the other. Hence we have two perspectives, both necessary for describing technical artifacts adequately, but we do not know how they are related even though there is no denying that they are.

### 1.3 Problems and Questions

The closing observation of the previous section brings us to the connecting theme of this thesis: the relation between the intentional and physical perspective as exemplified by artifacts' having technical functions and their being physical objects. My goal is to shed some light on this relation by homing in on two contexts in which technical functions and physical structures are

regularly connected in a systematic fashion. First, that of engineering design processes and, second, that of explanation, in particular explanations of how an artifact fulfills its technical function. To get the central questions into focus I will elaborate a bit further on both contexts.

Design processes are complex activities in which many different things happen. Notwithstanding this complexity it is generally conceded that the core of technical design is plausibly reconstructed as a systematic transition from a client's goal to a detailed description of an artifact plus instructions on how to use this artifact to achieve the goal (cf., for instance, Asimow 1962; French 1985; Gero 1990; Roozenburg and Eekels 1995; Pahl and Beitz 1996; Eggert 2005). This means that, in general, both the starting point and the endpoint of a design process consist of intentional descriptions. Design starts with goals or desired functions and it ends with a description of an artifact with a function and use instructions. Nonetheless, an element of this larger transition is a transition from a (desired) technical function to a physical structure. To see that consider the following. To come up with an artifact that can be used to achieve the client's goal, a designer has to devise a series of actions that are to be undertaken with the prospective artifact so that the goal will be realized. What the artifact is supposed to do in response to these actions is what will become its function. The challenge for a designer is then to draft a description of a physical structure that will actually behave as intended – in other words, to find an artifact that can fulfill this technical function. This smaller transition from a desired technical function to a physical structure is a transition from the intentional perspective to the physical perspective. It is this transition that will take center stage in the following chapters.

As our artifact-packed world testifies, designers are immensely successful at making this function-structure transition. What is more, they do so in systematic ways; engineering design is a broadly rational process. If it were the case that designing engineers are just really lucky every time they connect function and structure in the right way, or possess a remarkable but inscrutable creative gift that equips them for their job, there would be nothing much to gather from the study of design processes. Of course there is plenty of evidence that speaks against these suggestions.[3] Engineers and designers receive years of formal training at modern universities, in which they are taught a lot of basic and applied science. Design methodology, the field that aims to describe and develop systematic design methods, is a thriving academic discipline. By way of an *ad absurdum* argument, I reckon most engineers would take it as an insult if you told them they were just lucky or nothing but creative. They would

[3] Please note that I am not denying creativity its rightful place in design. I am only saying creativity is not the entire story; rationality also has a big part to play.

retort that design requires rational decision-making, good reasons, and hard analytical thinking. It is, at any rate, no less systematic and rational than other problem solving processes in science or everyday life.

A design process, then, is among other things a systematic way to get from a desired technical function to a description of an artifact's physical structure. A careful study of design processes and the tools, methods, and techniques used in them is therefore highly likely to reveal something about how designers are able to connect function and structure and thereby yield clues as to how the intentional and physical perspective are related. This task will occupy us in the next three chapters.

The second context in which technical functions and physical structures are brought in contact is that of explanation. The engineering disciplines presumably produce many sorts of explanations, but among them certainly are those that explain an artifact's technical function on the basis of its physical structure. As the 'author' of a technical artifact, a designer will be able to explain how the artifact can perform its function. If not, that would be serious reason to doubt her competence as a designer. The basic idea is this. Assume that an artifact has been designed successfully. It then has a technical function, which is to say – barring malfunction and other irregular cases for now – it has an ability to do certain things that help its users attain a practical goal if they use it in the proper fashion. For instance, a computer mouse can be used to operate a computer by holding it in your hand, moving it around, and clicking its buttons. It has this ability in virtue of its structure; its components in their particular configuration and their physicochemical make-up. Explanations that explicate how this is so are our second window on the connection between the intentional and physical perspective.

Designers, engineers, and technicians give such explanations when they describe why and how something works. After all, this involves giving information about the components of an artifact, what these components do, what they are made of, and how their material constitution underpins their behaviors. Sticking to the mouse example, an account of how the mouse works would include information about its conveniently shaped plastic casing, the little red light on its underside, how the light signal is converted into an electronic signal that is sent to the computer, the buttons and how clicking them generates another electronic signal, and so on and so forth.

Since we agreed that technical function is a notion from the intentional perspective, this cannot be the entire story. Information about an artifact's physical structure will never suffice to fully explain its technical function. For that, we also need information about the artifact's intentional and social context. Hence, the purely physical 'how does this work' story must

be supplemented with or modified by a further story that accounts for the intentionality inherent in an artifact's having a technical function. In chapters 5 through 7 I work out the details of this explanatory connection.

A this point the reader may well wonder what, if anything, the relation is between the two contexts in which the intentional and physical perspectives are connected. Do design processes and the explanation of artifacts have anything to do with one another? And does an account of how the design process gets from function to structure provide building blocks for an account of artifact explanation? Yes and yes. It would be odd indeed if the designer of an artifact would not be able to also give an explanatory account of how the artifact she designed works. This must mean that the knowledge required for a successful design process also provides the epistemic means to construct an explanation of how the resulting design works. Why should that be so? Because, I will propose in chapter 6, the explanations we are looking at are descriptions of the artifact's components and their configuration, behaviors, interactions, and physicochemical make-up. As a designer you would be the one who selected these components, thought about their precise configuration and what they would do to produce the required behaviors. That, as you will recall, was at the core of the design process: finding a description of a physical structure that can perform the desired function and so meet the client's needs. Hence, going through a successful design process almost guarantees that you will end up with the knowledge required to construct an explanation of how your designed artifact works. (Putting aside occasional false beliefs a designer may have about an artifact's and its components' dispositions.) Put more abstractly: If you move systematically from function to structure, you are guaranteed to also be able to account for function in terms of structure. Perhaps a brief comparison with science can clarify this point further. Unlike designers, scientists are not in a position of privileged epistemic access to their subject. They never get to choose the parts and components of the things they are studying. Scientists have to unearth the ontology of their subject area and, naturally, they may err at that. In contrast, designers get to fix their own ontology, so to speak. This gives them foolproof knowledge of a number of explanatorily relevant issues, viz., an artifact's components, their configuration, and their material constitution. They will often also possess correct explanations of all behaviors of the artifact and its components, but this is not automatically guaranteed.[4]

We will come across a specific instance of this general phenomenon later

[4] Designers can be mistaken about the workings and interactions of components. The steam engine apparently was an example of this. Although designers were reasonably successful in building working steam engines, they had a flawed understanding of why and how they functioned (Hills 1989).

in this thesis, when we explore the double role of functional decomposition. Functional decomposition is a method recommended by design methodologists. It consists in breaking down the desired function of an artifact into simpler sub-functions. As such, it figures as a method for making part of the transition from function to structure in the design process (chapter 4). The product of this method however, *a* functional decomposition, is part of an explanation of how the structure of an artifact enables it to fulfill its function (chapter 6).

The key issues for this thesis are now on the table. The connecting theme is the relation between the intentional and physical perspective on artifacts as exemplified by the relation between an artifact's having a technical function and its having a particular physical structure. I approach this theme by scrutinizing two contexts in which these perspectives are regularly and systematically connected: the design process and the explanation of artifacts. In sum, the two guiding questions are as follows:

(1) How does the design process realize a transition from a description of a technical function to a blueprint for an artifact's structure?
(2) How are an artifact's physical structure and its technical function connected in explanations of how an artifact performs its technical function?

## 1.4 Background

In this section I want to put the project of this thesis in perspective by relating it to recent developments in the field of philosophy of technology. Although this thesis is a work in the philosophy of technology, it is not so in the sense in which that field has traditionally been conceived. Traditional philosophy of technology, as exemplified in the work of notable philosophers like Karl Jaspers, Martin Heidegger, and Jacques Ellul, has tended to focus not so much on technology itself, but on the effects of technology on individuals and society at large or on preconditions of technology and technological development, such as a technological 'mode of thinking' or even 'way of being'. This focus has made up the bulk of philosophy of technology for quite some time (cf. Schuurman 1972, 1990; Ferré 1988; Volti 1988; Mitcham 1994). The outcome of such philosophizing was more often than not a pessimistic diagnosis; technology was perceived to be the result of a (morally or otherwise) reproachable way of being in the world (Heidegger), to threaten the autonomy and uniqueness of individuals (Ellul, Jaspers), to alienate people from their material environment (Jaspers), to lead to and reinforce a defective and depraved worldview (Heidegger, Jaspers), or to possess a dangerous developmental dynamic of its own (Ellul). Joe Pitt's impression is that philosophy of technology has largely

been 'social criticism ..., critical denunciations of the negative effects of technology on human values and the human life' (Pitt 1999: vii).

No doubt the questions of traditional philosophy of technology deserve the philosophical attention they got and the philosophers mentioned have done a great job in addressing them. Nonetheless, skimming through their writings one wonders whether their thinking has at times not been a bit too lofty, far-removed from the nitty-gritty of actual technology. For one thing, it seems unlikely that one could adequately characterize the essence of a vastly multifarious phenomenon like modern technology as one particular mindset, or evaluate its effects on contemporary society in one fell swoop. Yet this seems to be what many traditional philosophers of technology have sought to do. However, this is not the place to discuss the merits of traditional philosophy of technology – nor am I particularly qualified to carry out such a discussion. I merely want to note the fact that traditional philosophy of technology has not engaged in extensive empirical study of actual technology. By and large, it has taken an *externalistic* approach to technology, looking at it from the outside and treating technology itself as a black box (Kroes 2000: 22). Regardless of whether one thinks this is good or bad, one can at least agree that the field may benefit from philosophical projects that do incorporate such empirical material. I am not original in this suggestion. In fact, it is more accurate to say that I am merely repeating the call of some of my better-informed predecessors. Carl Mitcham's *Thinking Through Technology* asserts that philosophers should pay more attention 'to what really goes on in engineering and technology' (1994: ix); the proper way to do philosophy of technology is 'in a way that does not exclude engineering discourse' (ibid.: 267). In a similar vein, Joe Pitt urges that epistemological and ontological questions about technology ought to take logical priority over social criticism and other evaluative projects. 'Understanding what we know about technology, and understanding *how we know that what we know is reliable*, are the prerequisites to offering sound evaluation of the effects of technologies and technological innovations on our world and lives (Pitt 1999: viii).

It is out of these sentiments that the Dual Nature research program arose (cf. Kroes and Meijers 2000). To motivate the orientation of this thesis, I will summarize the guiding ideas behind this program, especially those that relate to the proper way of doing philosophy of technology. In the introduction to the volume that can be considered as a founding document of the program, and perhaps even of a new style of philosophy of technology, Peter Kroes and Anthonie Meijers argue that philosophical reflection on technology 'has to be informed by what goes on 'inside' technology; it has to be based on an understanding of technological practice itself, of how technological artifacts are designed, developed, and produced by engineers' (ibid.: xviii). They recommend that the philosophy of technology focus more on descriptive and non-moral

topics, rather than the traditional normative and moral topics. This means that, instead of morally evaluating technology and its effects, it should start by looking at what really goes on in the design, development, manufacture, and maintenance of technologies, at the knowledge used and developed in technology and engineering, at the sort of actions that characterize technological practices, etc. It does not mean that philosophy of technology should transform itself into a branch of empirical science; its business continues to be conceptual issues. In particular, 'the clarification of basic concepts and conceptual frameworks employed in scientifically adequate descriptions of parts or aspects of technology' (ibid.: xxiv).

In his contribution to the same volume, Peter Kroes elaborates on this theme by proposing a threefold reorientation for philosophy of technology (Kroes 2000: 27–28). First, it should shift its focus away from the use phase of technologies to the design, development, and production phase. Not only does this phase pose philosophical questions that are interesting in their own right, but also is analysis of the design phase likely to facilitate later treatment of philosophical issues regarding the use phase. Second, the level of analysis should change from highly abstract and global to concrete and local. Instead of lumping all forms of technology and technologies together, philosophers of technology should be sensitive to the differences between technologies, technological practices, and engineering disciplines. To assume that technology is a monolithic entity and that all its manifestations share a common core is simply unwarranted. Third, empirical material ought to play a more important role in philosophy of technology. Philosophical analyses ought to be backed up by studies of what actually goes on in technology.

The following chapters constitute a modest attempt to heed the call to an empirical turn in the philosophy of technology. You will not find moral evaluations of technology in them, nor will they be about detrimental effects of technology on individuals or society. Instead, the focus is on analyzing and clarifying what goes on in design processes and adequately understanding explanations engineers produce. The scope of my work is thus limited to the design phase, as opposed to the use phase. It is also limited in that it does not purport to cover all types of design processes, but only a limited subclass. Roughly, I will limit myself to the design and explanation of material artifacts that are used by people for practical purposes. I specify this limitation further in the next section. Finally, the starting point for my philosophical explorations will be empirical material: the work of design methodologists on design processes and methods and tools used therein (chapters 3 and 4) and actual explanations given by engineers (chapter 6).

In closing this section I should add that, along with a more empirical stance, the Dual Nature program also adopted the analytic style of philosophizing. This

coupling is contingent. More empirically informed philosophy of technology need not also be analytic. In fact, those favoring a European continental style of doing philosophy have voiced many the same concerns about traditional philosophy of technology and issued similar calls for an empirical turn. They, too, have embarked on projects to remedy the lopsided development of philosophy of technology. To mention but one example, Peter-Paul Verbeek's *What Things Do* (2005) insists on the need to take technology itself more seriously and presents detailed investigations of how technical artifacts alter the ways in which we perceive and deal with the world.

### 1.5 Some Limitations

I want to use this section to mention some limitations on the sort of artifacts, design processes and explanations that I intend to study. For the average person reading these lines a quick glance around the room in which she sits should suffice to demonstrate that the world we live in is densely populated with the products of technology. So much so, in fact, that as I am writing these pages I have a hard time discovering something that is not in any sense an outflow of human efforts to make the world better suited to our likings. A lot of the things I see are *technical artifacts*, material objects that are used by people to realize practical goals. These products of technology take center stage in the following chapters.

The 'technical' in 'technical artifacts' serves first of all to demarcate the artifacts I am interested in from other types of artifacts, such as social artifacts (money, marriages, organizations, etc.), aesthetic artifacts (paintings, sculptures, music, and other works of art), non-intentional byproducts of human activity (saw dust, footprints, etc.), and artifacts in the sense of undesirable distortions of measurements and experiments that are the result of the specific experimental setup. Even with these types of artifacts barred, technical artifacts come in incredibly many varieties: software code, satellites, skyscrapers, scissors, automated production lines, prostheses, and nanorobots. Such artifacts have little in common, and their design and explanation may well look very different. I will therefore limit my attention to a subclass. Given the rich variety in which artifacts come it is practically impossible to give precise limits, but I will try to be as specific as I can. Roughly, the things I have in mind as paradigm cases are medium-sized concrete material objects that people use for specific practical purposes. They are the sort of artifacts designed by mechanical engineers, electrical engineers, civil engineers (their objects often being bigger than medium-sized), aerospace engineers, and a lot of industrial designers. Excluded are artifacts that are primarily designed for their aesthetic qualities (obviously works of art, but also some architectural projects and products of industrial design). What I have to say might still apply

to the 'technical' aspects of the design of such artifacts, but that will not be a touchstone for my analyses. Also excluded are immaterial artifacts, such as software, algorithms, models, rules, and other abstract entities. Finally, I want to exclude general-purpose artifacts and stuffs (chemical substances, flour, steel, etc.), which lack specific functions. Presumably, a lot of materials engineering and (bio)chemical engineering is about such technical artifacts. Again, what I have to say might apply to them as well but that will not be a litmus test.

Another limitation lies in the aspects of design that I take into account. I zero in on the function-structure transition. Some will perhaps feel that this is too narrow-minded. After all, design is embedded in a socioeconomic context where all sorts of non-technical factors influence its outcome (for elaborate examples, see Bijker *et al*. 1987; Bijker 1995; Pinch and Oudshoorn 2003). It is often carried out by teams of designers from different backgrounds, leading to confusion, conflict, or even more fundamental communication problems. Larry Bucciarelli (1996) even goes so far as to suggest that sometimes members of design teams live in different 'object worlds'. All of this is true and I cannot but encourage that everything related to such factors be studied in great detail by the cooperative efforts of design methodologists, psychologists, sociologists, anthropologists, philosophers, and whoever else has something to contribute. That, however, is not the task I have set for myself. I choose to ignore these influences and focus exclusively on the technical aspects of design. In doing so, I implicitly assume that all the factors I mentioned (and possibly others I am not aware of) merely complicate design processes but do not fundamentally alter their nature as a transition from function to structure.

### 1.6 A Bit of Methodology

I do not pretend to have anything particularly original or surprising to say about the methodology for my project. I feel comforted, however, by the fact that I share this feature – should I say vice? – with most contemporary analytic philosophers. Stepmotherly treatment of philosophical methodology seems to be accepted practice.

First and most importantly I stick to the methodological canons of contemporary analytic philosophy. Roughly that means concepts, analysis, and argument are the raw material with which I work. I start from concepts that play a role in everyday language and engineering discourse and try to clarify and, if necessary, precisify them by means of analysis supported by rigorous argument. As can be glanced from the above, key concepts in this thesis are artifact, function, structure, design, and explanation. The goal is to elucidate these concepts in a way that 'makes best sense' (Jackson 1998: 36) of how people, especially engineers, use them. Sometimes, this may require

some reconstructing and reinterpreting of what people say for the benefit of increased theoretical perspicuity. How much revision is allowed is hard to say in advance and in general terms. I think sound judgments can only be made on a case-by-case basis, depending on such factors as the interpretational latitude of ordinary locutions, the centrality of the concepts to our understanding of the world, the alleged size of the theoretical benefits, and plain old common sense. Even then, presumably, matters of personal taste will continue to influence our decisions as to which sort of revisions are legitimate. What some will think of as an unacceptable departure from our everyday use of a concept, others will take to be but a small sacrifice in the interest of increased clarity, or perhaps even a healthy correction of confusions lingering in ordinary language. By way of a general maxim, I can say that I will try to be as faithful to everyday and engineering discourse as can be. I will flag any proposed deviations as being deviations.

Second, in line with Quine's famous critique of the analytic-synthetic distinction (Quine 1951) and the general outlook of the Dual Nature program, I sympathize with the view that philosophy and empirical science are not entirely disparate enterprises. Seeing them as largely continuous with one another should stimulate interaction, which can only benefit both sides. For this reason I feel bound to take empirical data seriously. For my particular situation, this implies that engineering discourse and practice has to be the starting point and touchstone for my analyses. For information on design processes, I turn to design methodologists and for real examples of artifact explanations, I look at patents and engineering handbooks.

The choice to look at design methodology, as opposed to directly at design processes themselves is motivated by the following considerations. First, efficiency. The sheer time it would take to study a representative batch of design processes far exceeds the time I had available. Design methodology has been studying design processes for decades and thus provides a convenient shortcut for a relative outsider like myself who wants to learn more about design processes. Second, I am not so much interested in quirky details of actual design processes, but in the centrally important activities and steps that make up the essence, so to speak, of design. Actual design processes may exhibit various distortions due to idiosyncrasies of the participating engineers or special circumstances. One of the goals of design methodology is to find just those essential activities and steps. So rather than trying to weed out the contingent distortions and irrelevancies myself, I may just as well use the filter of design methodology. Third, I am not interested in design processes per se; I want to know how the transition from function to structure is accomplished in *successful* design processes. The goal of design methodology is of course to describe successful design processes and improve them further.

The general question of how empirical material can or must influence philosophical analysis remains. On the one hand, it is typically not the case that the empirical data necessitate a single philosophical account. If, as the Quine-Duhem thesis famously asserts, underdetermination besets scientific theories, then it besets philosophical theories *a fortiori* since they typically reside at an even higher level of conceptual abstraction than their scientific counterparts. On the other hand, I have been championing a greater role for empirical data in the philosophy of technology. To reconcile these two points we must realize that philosophy is not an empirical science and that its theories will never be fully dictated by the empirical data. But that does not mean we must fall into the other extreme of neglecting the empirical data where it does point in clear directions. By beginning with engineering practice and constantly keeping in touch with it, I try to render justice to the empirical data as far as possible for philosophical analyses. Again, I have no general guidelines as to how this should be done. I do offer some more specific methodological remarks for my special case in section 3.6, where I sort out how a philosophical account of technical design can engage with the findings of design methodology.

### 1.7 Outline and Reading Guide

To finish the chapter, I will provide an overview of what is coming and some reading suggestions for different possible audiences.

Chapters 2 through 4 focus on the first context in which function and structure are systematically connected: the engineering design process. Chapter 2 fills the theoretical and conceptual toolkit for this thesis. In it, I present a more elaborate philosophical account of design and a function theory developed by my colleagues. I also define a number of concepts and the distinctions between them. With these concepts the central notions of technical function and physical structure can be analyzed more rigorously. Application of the above theoretical tools yields the conclusion that my question about design actually involves three different issues. Of these three, one is relatively easily resolved but the two others merit further attention.

Since the guiding question is how engineers get from function to structure, chapter 3 surveys what design methodologists say about design processes to see if that contains any direct hints as to how this transition is made. It will turn out that the design methodologists I study are not very sensitive to the distinction between function and structure. As a result, their work contains little that is directly relevant to my problem, although it does contain a lot of useful information about methods and tools that aid designers in their job.

These methods and tools are the key to the indirect route I explore next. If design methodologists' descriptions of design processes do not make clear how the transition from function to structure happens, then a closer look at their

practical counsel might reveal more about the transition. Assuming that the methods and tools recommended by design methodologists tend to be effective, a study of how these methods and tools work must reveal something about how they help in getting from function to structure. In chapter 4, therefore, I scrutinize one such method: functional decomposition. By drawing on philosophical accounts of function, behaviors, and dispositions I clarify what functional decomposition consists in. It turns out that design methodologists use the name 'functional decomposition' for quite a few different things. Unraveling these meanings will take up a large portion of the chapter, but we will also have the opportunity to explicate how the different activities that go under the name of functional decomposition facilitate part of the transition from function to structure.

Chapters 5, 6, and 7 home in on the second context in which function and structure are brought into contact; the explanation of artifacts, or, more accurately, the explanation of how an artifact fulfills its technical function in virtue of its physical structure. Explanation has been a centerpiece of extensive philosophical debates in the past six or so decades. To position my own work, I will use chapter 5 to outline the available philosophical theories of explanation.

Chapter 6, then, is about artifact explanation. I start with a couple of examples of explanations given by engineers and designers. Analysis of these examples gives rise to an account of explanations that account for an artifact's overall behavior in terms of its sub-behaviors and the physicochemical make-up of the artifact and its components. This account of artifact explanation builds on the well-known causal-mechanical model of explanation and some recent amendments to this model.

Since technical functions presuppose intentionality, the account of artifact explanation from the previous chapter must be supplemented with a further account that puts the intentionality back into explanations of technical functions. There are good and bad ways to do this. In chapter 7 I discuss and reject what I think is a bad attempt. My diagnosis of what is wrong with this attempt will also provide the building blocks for my own positive proposal. I argue that the intentionality and normativity that surround technical functions must be accounted for by a function theory, i.e., a theory that gives the necessary and sufficient conditions for when artifacts have functions or when ascriptions of technical functions are true or justified. This function theory can then be superimposed on an account of artifact explanation. On the resulting picture, connecting an artifact's physical structure with its technical function is a two-tiered affair: A function theory specifies the social-intentional relations that pick out an intended behavior as an artifact's technical function. Artifact explanations can then be used to explain how an artifact is supposed to be able to show this intended behavior. If an artifact functions properly, this

explanation is at the same time an explanation of how the artifact actually has the disposition to show the behavior that is its function. If it malfunctions, an artifact explanation explains just how the artifact's designer or users intended or believed it to have the disposition that is its function.

Chapter 8 sums up my findings and provides a final look at how the technical function and physical structure of artifacts and the intentional and physical perspective on artifacts are related. I also point out avenues for further research. Furthermore, I look back at the initial Dual Nature intuition and reformulate it with the help of my more precise conceptual framework. And I show how my arguments in chapter 7 naturally lead to a compelling account of artifact kinds that has recently been proposed in the literature.

Finally some reading advice: I think the primary audience, to wit philosophers of technology and philosophers of science, should find all the chapters worthwhile. They may choose to skip chapter 5 if they are cognizant of the theories of explanation on offer. Other philosophers, if they are interested in functions, events, or explanation, should try chapters 4, 6, and 7 to see if they contain anything of their liking. Philosophically minded design methodologists and designers would do well to focus on chapters 2 through 4 and might want to skim through chapter 6. My advice for non-philosophers who have gotten this far is to move directly to the concluding chapter to pick up the overall line of argument of this thesis. If there is anything in particular that you want to know more about, the table of contents can then guide you to it.

# Design, Function, and Other Key Notions

## 2.1 Introduction

This chapter serves to clarify some centrally important notions from the previous chapter further and to introduce and define a few further concepts. It is largely expository, presenting and summarizing relevant work by others. With the resulting theoretical toolkit I can show that the central questions as I formulated them in the previous chapter actually contain three different issues, which are better taken on separately.

In section 2.2 I start by placing the brief and somewhat provisional account of design processes from the previous chapter in the context of the philosophically more developed action-theoretical account of design developed by my Dual Nature colleagues. This allows me to delineate the specific focus of my project more precisely. Next, in section 2.3, I present my assumptions about technical functions, again with the help of an existing function theory. In section 2.4 I get back to the central questions from the previous chapter and analyze them as getting at three different issues. To formulate these three issues precisely, we will need the distinction between dispositional and categorical descriptions. I present Stephen Mumford's account of that distinction in some detail in section 2.5. I conclude this chapter in section 2.6 by reformulating the main questions for my project.

## 2.2 A Philosophical Account of Design

Let me start by briefly repeating what I said in the previous chapter. I characterized a design process as a transition from a client's goal to a blueprint for an artifact plus instruction on how to use it. As part of this larger transition, designers have to get from a desired technical function to a description of the physical structure of a prospective artifact. Only this latter transition is a transition from the intentional perspective to the physical perspective on artifacts and that was part of the reason I decided to focus on it.

I now want to take another look at the broader process of design to both clarify and motivate my focus a bit further. Some colleagues of mine developed what I think is a compelling action-theoretical account of design processes (see Houkes *et al.* 2002; Houkes and Vermaas 2004, 2006). The key tenet of this account is that engineering design is both the design of *use plans* and the design

of technical artifacts. Roughly, this means that design is just as much about coming up with considered series of actions by which some goal can be achieved as it is about creating an artifact that can be used to perform these actions. Artifacts are thus always embedded in use plans.

The notion of a plan is borrowed from general action theory, where it denotes a considered series of actions, undertaken with the purpose of achieving some goal. A use plan, then, is a considered series of actions, undertaken for achieving some practical goal, where at least one of the actions in the series involves the manipulation of a technical artifact. The central claim of the account is that design, characterized as a transition from a client's need to a description of an artifact's structure, is best understood as a two-tiered process: It starts with the design of a use plan and, if necessary, is complemented with the design of one or more artifacts.[5] The first tier is *use plan design*: Designers take a client's need or goal and conceive a use plan that will lead to the realization of that goal. Some of the actions in this use plan, at least one, involve the use of at least one artifact. The described use of an artifact specifies, often in a fairly abstract way, the behavior that an artifact must exhibit: If it is used in such-and-such ways, it must do so-and-so. It is this behavior that becomes the artifact's technical function if the design process is successful and the artifact is put to use according to the designer's intentions. If the used artifact does not yet exist the second tier, *artifact design*, is initiated. Designers start with the desired behavior as specified in the use plan and devise a detailed structural description of an artifact – a blueprint – from which a complete specification of its configuration and constituting materials and components can be derived. The final output of a design process is both (a description of) an artifact with a specific technical function and a use plan for that artifact. Houkes *et al*. (2002) and Houkes & Vermaas (2006) give detailed descriptions of both design types, but there is no need to rehearse those here. The interested reader can consult those sources herself.

In everyday language it is common to use 'design' in a narrower sense, exclusively for artifact design, thus taking use plan design for granted. This usage is perfectly natural since many cases of design concern improved or similar versions of an already existing product so that the use plan can indeed be taken for granted. But this does not imply that the design of use plans should not be part of a general account of design; it only means that in real-life circumstances use plan design can often be skipped because designers rely on existing use plans. As I intend to focus on artifact design here I, too, will use 'design' primarily to refer to artifact design.

[5] These two things do not typically take place in chronological order; they rather co-evolve through synchronous iterative cycles. They are only distinguished for analytical reasons.

As to the status of this account of design: It is a *rational reconstruction* of design processes. Negatively, this means it is neither supposed to be an accurate description of what actually goes on in design, nor a prescriptive account of what designers ought to do. Positively, it means that although the account is based on descriptive models of design, it also adds elements to these descriptive models – most prominently use plans – and it imposes a norm of rationality broadly conceived. Since it is a rational *reconstruction*, it is not committed to the claim that every design process is in fact carried out as the reconstruction says or that designers actually use the terminology of the reconstruction in their own thinking and deliberating. Nobody requires that they, say, talk explicitly about use plans. Real designers may take very different steps, describe their work in completely different terms, or even be highly irrational. The account is only committed to the weaker claim that actual design processes, in so far as they are rational, can be recast in its terms. (See Vermaas & Dorst (2006) for this difference between rational reconstructions and descriptive and prescriptive models of design.) My own work on design, especially in chapter 4, is in line with the status of this model; it is also supposed to be a rational reconstruction of things that happen in design.

Without doubt this rational reconstruction leaves out massive amounts of detail about what actually happens in design processes. But then again, that is exactly the point of a reconstruction. To mention but a few obvious things, the starting point of a design process will frequently not be a description of only behavior; more often than not there are physical constraints and specifications right from the start. Analogously, the endpoint may still contain descriptions of behavior and functions, for instance when only the technical function of standard components or parts is given because there is no need to specify the structure of standard components. Further, many design processes consist in optimization of an existing product or cosmetic improvement of previous versions of an artifact. In such cases, most of the structure is already given at the outset and there is hardly any translating left to do. Next, design assignments are often refined and adjusted through several iterative steps, based on changing needs of the client, economic and financial factors, perceived physical impossibilities, social conventions, legal or other rules, etc. All such things, and many more, are enormously important in actual design processes and must be studied if one wants a fully accurate and detailed picture of design. My project in this thesis, however, is different. I want to know how designing engineers get from function to structure. While the abovementioned factors may cloud a direct view on this transition, they do not detract from the fact that this transition happens in many design processes and that therefore design processes are a rich source of information about how this transition can be made.

My question, then, is about the transition from desired behavior to a physical structure. In terms of the above account of design that is essentially artifact design, and not use plan design. Hence, I will not talk about how designers cook up plans to achieve goals. I assume these use plans as given. This is not because I think the creation of use plans is unimportant or trivial. Use plan design is a topic right in the heart of philosophical action theory. It has plenty of intrinsic value apart from any instrumental value it may have for an account of engineering design. Planning is probably one of the most distinctive features of the human cognitive enterprise. Its relevance and importance far exceed the confines of philosophy of technology and engineering design. The choice to not include it in my project is motivated by the need to narrow my project down to manageable proportions and emphatically not by a mistaken (negative) value judgment about its philosophical interest. (And, of course, use plan design does not bridge the gap between the intentional and physical perspective, since goals and actions are all intentional notions. This, as the reader will recall, was the reason to concentrate on the function-structure transition in the first place.)

Let me point out a few consequences of this limitation. First, by not including the transition from a goal to a use plan, I loose sight of how a goal is translated into actions and how designers deal with the underdetermination of plans by goals. People may consider this unfortunate but, as I said, it is a topic far too big to be discussed only in the context of engineering design. It would make for another book-length project in its own right. Second, by assuming use plans as given I also leave out part of the process in which the 'division of labor' between artifact and user is determined. Lots of choices must be made in design processes about what exactly a user does and what the artifact does: Push an on/off button or plug in and unplug the device; rotate the screwdriver by manual force or turn a dial to activate an electromotor; let the user feed both sides of a sheet paper of paper into a printer manually or have an inbuilt module that turns the paper over for double-sided printing; etc. Grounds for such decisions presumably include multifarious considerations: cost, time constraints, efficiency, effectiveness, client desires, feasibility, etc. I assume that the outcomes of those decisions have been taken into account in the specification of the artifact's desired behavior(s). I do say a little something in chapter 4 about how functional decomposition specifies a brief initial description of desired behavior in much more detail. This specification is most likely the outcome of a complicated decision process of the sort I just hinted at, but of which I will speak no more.

Perhaps the reader wonders what it is that I do take into account then. The answer to this question must wait until after the next section when we have a clearer understanding of what technical functions are.

## 2.3 Assumptions About Technical Functions

The next key notion for my project is that of a technical function. So far, I have been using the notion in an intuitive sense but now I want to make it more precise. I do so, again, by drawing on material developed by colleagues in the Dual Nature program who have done a lot of work on analyzing this notion and differentiating it from that of a biological function (see Vermaas and Houkes 2003, 2006; Houkes and Vermaas 2004, 2006). For most of my project I do not need a comprehensive account of technical functions, so I will start with some assumptions about the notion of function that I do need. After that, I will present the ICE-theory of technical functions developed by Wybo Houkes and Pieter Vermaas as an example of a more developed account of technical functions that could suit my project rather well.

I assume that a technical function is the behavior or disposition[6] for which an artifact was created by its designer or selected by a (later) community of users. Since artifacts are always embedded in use plans, we may also say that an artifact's technical function is the role it is supposed to play in a socially accepted use plan, which might issue from the artifact's original designer or from a community of users. Having a technical function is thus not an intrinsic property of an artifact, but a relational property with as *relata* the artifact and some community of intentional agents (see also Meijers (2000), Vermaas (2006) and Scheele (2005) for this relational nature of functions).

Mere intentions do not suffice to bestow functions upon artifacts. The beliefs on which the intentions are based must be of decent quality, so to speak. A designer's beliefs about the materials and components she is working with, their behaviors and interactions, and hence about what the artifact can do have to have positive epistemic status, i.e., be justified, warranted, or what else is deemed appropriate.[7] In addition, the designer's intentions must be carried out reasonably successfully, which is to say the actual artifact must a good approximation of what the designer had in mind. As a shorthand for these two things, we could say a designer has to be *competent* in order to bestow the artifacts she designs with functions.[8] I take it that artifacts sometimes change

[6] Dispositions and behaviors keep cropping up. For now it suffices to see that the two are closely related: A disposition manifestation is a behavior and having a disposition is being able to behave in certain ways. A disposition is always a disposition to *do* certain things. In section 2.5 I offer a detailed picture of the connection.

[7] What the necessary and sufficient conditions are for beliefs to have positive epistemic status is a central question in epistemology. Nothing here depends on a specific answer to this question.

[8] To bar other funny cases, we may have to add that a designer has to be *benign*, i.e., set on actually satisfying a client's need as opposed to intentionally creating something that has another function or does not function at all.

their technical function when they are co-opted into a new standard use by some community of users. For such changes in function to occur, the users, too, must have beliefs with positive epistemic status about what the artifact can do. Cell phone users cannot collectively decide by fiat that the new function of cell phones is tele-transportation, because they do not (and cannot) have beliefs with positive epistemic status that cell phones can fulfill this function. Note that this approach does not exclude the possibility of malfunction; beliefs with positive epistemic status may turn out to be false, artifacts may break down after a period of use, or they may be meddled with and so be rendered dysfunctional.

These are my assumptions about artifact functions. They clearly do not add up to a mature account of technical functions, nor do they give precise conditions for when artifacts have function, but they do point a way to those things. Presumably some fine-tuning is necessary to deal with atypical cases. Think of a lonely designer who devises a contraption but fails to communicate its use plan and function to anyone. Does this thing nonetheless have a technical function or not? When do artifacts lose their function? It seems reasonable to say a blunt knife still has its function, but a completely malformed piece of metal that used to be the blade of a knife probably does not. I think there is no sharp cut-off point to decide such cases, but it may be convenient to build one into a function theory to prevent implausible claims about heaps of scrap and the like having functions. I leave such challenges for others to sort out and will now turn to a specific example of a function theory that fits my assumptions well enough to make it plausible that a precise function theory along these lines can be developed. Readers who are willing to take my word for this may skip directly to the next section.

The ICE-theory of technical function by Wybo Houkes and Pieter Vermaas is created out of three simplistic accounts of artifact functions: the Intentional theory which equates functions with the unqualified intentions of designers or users; the Causal theory which says that functions are any causal role an artifact might play; and the Evolutionary theory that identifies functions as that for which the artifact's predecessors were selected for. I do not think the theory is accurate as it stands, but there is no need to detail my quarrels here. We will get to that in chapter 7 when we take a closer look at the theory.

The basic idea of the ICE-theory is to give the conditions under which an artifact's intended disposition can be ascribed to it as its technical function. Roughly, a technical function is that disposition that, when manifested, contributes to the goal of a use plan. Technical functions encompass both accidental and proper functions (see Millikan (1984; 1993) for one of the most sophisticated discussions of this distinction for biological functions). The distinction between them is drawn by means of the social-epistemic distinction

between professional designers and laypersons. When a use plan issues from a professional designer or engineer and the use is socially accepted, the technical function is a proper function. If not, it is an accidental function.

Here are the official conditions:[9]

> An agent *a* ascribes the capacity to $\phi$ as a function to an artefact *x*, relative to a use plan *p* for *x* and relative to an account *A*, iff:
> (I) The agent *a* has the capacity belief that *x* has the capacity to $\phi$, when manipulated in the execution of *p*, and the agent *a* has the contribution belief that if this execution of *p* leads successfully to its goals, this success is due, in part, to *x*'s capacity to $\phi$;
> (C) The agent a can justify these two beliefs on the basis of *A*; and
> (E) The agents *d* who developed *p* have intentionally selected *x* for the capacity to $\phi$ and have intentionally communicated *p* to other agents *u*.
> (Vermaas and Houkes 2006: 9)

A few points deserve further emphasis. First, a minor technical glitch. I think Vermaas and Houkes must mean something like: 'An agent *a* is justified in ascribing the capacity to $\phi$ as a function to an artifact *x*.' After all, the point of their theory cannot be to record what happens when agents ascribe functions to artifacts, because actual function ascriptions may be made for all sorts of mistaken or odd reasons. As a purely descriptive account of actual function ascriptions it is plainly false. Besides, there would be little philosophical value in the attempt to capture the psychological dynamics of function ascriptions. The theory should rather be seen as an account of when an agent is correct or justified in ascribing a function. Second, ontologically speaking an artifact's technical function is its intended disposition, i.e., the behavior it is supposed to be able to show. Not just any old disposition, but one that is singled out by means of the use plan in which the artifact is embedded and the social-epistemic context of designers and users that surrounds this use plan. This observation is in line with what I said in the previous section about design processes starting with desired behavior, which then goes on to become an artifact's technical function. Third, the ICE-theory is strictly speaking about function ascriptions by agents and not about the having of functions by artifacts. This is why the theory introduces the justified beliefs of the function-ascriber and does not directly mention the beliefs of the designer or users who bestowed the artifact with its function, as I did above. Fourth, justified or correct function ascription are relativized to a justificatory account *A*, which is itself subject to normal

[9] Note that Vermaas and Houkes prefer the term 'capacity' where I use 'disposition'. I take the two as synonymous.

standards of justification and acceptability. This account can take on a number of forms; for new, inexperienced users it can be simple testimony or observation (having heard or seen that this contraption is a laser pointer, or having read the inscriptions on the package), for technically savvy users who enjoy taking apart their electrical appliances, it can be practical insight in their internal workings combined with experiential knowledge, and for engineers, it will typically be full-fledged artifact explanations and scientific explanations, often combined with practical experience from prototype tests.

Even if one does not agree with all its details, this theory at least makes it plausible that a precise and adequate function theory can be developed along the lines of my assumptions in this section. And that was all I tried to argue here.

### 2.4 Another Look at Artifact Design and Explanation

Armed with the more precise understanding of the design process and technical functions I want to take another look at the central questions for this thesis. They were:

(1) How does the design process realize a transition from a description of a technical function to a blueprint for an artifact's structure?
(2) How are an artifact's physical structure and its technical function connected in explanations of how an artifact performs its technical function?

Let us revisit both questions in turn. As to the first question, in section 2.2 I pointed out what it aims to exclude from consideration. Let me now provide a positive characterization of the topics it does take into account. It zeroes in on the transition from desired behavior to a physical structure. Given the above account of design and functions, the exact nature of this transition looks as follows. The initial description of the desired behavior is derived from a use plan: It is the role of the artifact in the use plan. At least one of the actions in the use plan is an action that is carried out with an artifact. The description of this action specifies what the user is to do with the artifact so that it does certain things in response. In other words, the description of this action specifies the desired behavior of the artifact in response to specific inputs on the part of the user, which is to say it specifies a desired disposition or capacity of the prospective artifact. This behavior may be characterized in full detail but it need not be. At the very least a designer can derive from the use plan a rough characterization of what the artifact is supposed to do.[10] It is this desired or intended behavior that becomes the artifact's technical function if the design process is brought to a successful end.

Perhaps the reader has noticed how the present characterization of a design process as a transition from desired behavior to physical structure differs from the one used in the previous chapter and in the original Dual Nature program. Before, I said the starting point of a design process is a (desired) technical function; now I say it is desired behavior. The change has to do merely with precision. If you accept the above picture of design and want to be really precise in your wording, you should say a design process starts with desired artifact behavior. If the design process is carried out by a competent designer and is successful, this desired behavior becomes the new artifact's technical function. So technical functions are the outcome of design processes and not their starting point. Saying that a design process starts with a technical function is slightly anachronistic; it projects the successful outcome of the process back onto its starting point. In a loose sense, this is perfectly all right. Here is why: To learn about what happens in design processes, we should study successful exemplars in which the desired behavior indeed becomes the artifact's technical function. The assumption that a design process is successful is thus implicitly made and it follows from this assumption that the desired behavior automatically becomes the technical function. Because of that, the distinction between intended behavior and technical function is not all that important in the context of (successful) design.

The essential challenge in artifact design is thus to get from an initial description of desired behavior (desired disposition) to a description of a physical structure that will be able to carry out this behavior (has this disposition). This is what the first research question is about.

The notion of a technical function or that of intended behavior has several distinct connotations: A description of a desired technical function is first of all a description of *behavior*; it says something about what the artifact is supposed to *do*. Second, it is an *intentional* description in that technical functions and intended behavior presuppose intentionality. In addition, desired functions are typically described at relatively *high levels of abstraction*, i.e., not in the vernacular of physics and chemistry. A description of physical structure, on the other hand, has the following contrasting connotations: It is a description of an artifact in *categorical* terms; not about behavior or dispositions but about the material constitution that ultimately underlies these dispositions.[11] Second, it is a *physical*, i.e., non-intentional description of the artifact as a material object.

[10] I repeat that use plan design and artifact design are not to be understood as chronological phases of design. They are distinguished for analytical reasons. In reality use plan and desired artifact behavior may well co-evolve through several iterative cycles of specification. See also note 5.

[11] A precise account of the distinction between dispositional and categorical descriptions is given in the next section.

Third, it is implicitly taken for granted that this description is sufficiently *concrete* so that every detail about the artifact's structure is derivable from it (if it were not, it could not function as a blueprint from which the artifact is created). Given these three distinct contrasting connotations, the transition from intended behavior to a description of a physical structure in fact harbors three distinguishable components:[12]

(a) A transition from a description of *behavior* or *dispositions* to a description of *structure* or *categorical properties*;
(b) Usually, a transition from a fairly *abstract* description of an artifact's desired behavior to a very *concrete* description of the artifact's behavior and, ultimately, physical structure; and
(c) A transition from an *intentional* description of desired behavior to a *non-intentional* description of physical structure.

The key difficulty that is resolved in transitions (a) and (b) is that of *underdetermination*: the in-principle underdetermination of physical structure by behavior and the more practical type of underdetermination that results from the initial description of behavior's being abstract and imprecise. Let me explain. Any description of behavior, no matter how detailed and concrete, necessarily underdetermines the underlying physical structure because there are no *a priori* entailments from descriptions of behavior to descriptions of structure or categorical properties. Describing what something can do leaves open what its physical structure is. Even if we happen to know just one particular structure that has the described disposition that does not imply that there exist no other structures that have the same disposition. And even if, in actual fact, there exist no other structures with the same disposition, it is still conceivable that there would exist other structures. If only the initial descriptions of desired behavior with which design starts were very precise and formulated in terms of physics and chemistry, then this in-principle underdetermination problem might not be so bad, for there are of course plenty of *a posteriori* entailments from descriptions of behavior to descriptions of categorical properties that can underlie the desired behavior. We know by experience that if it boils at 100°C, it is probably $H_2O$ and if its electrical conductivity equals $0.0993 \times 10^6$/cm Ω, it is iron. This is not to say that design would be particularly easy in this case, for designers would still face complex decisions: Which of the possible structures is most suitable for design task

[12] These three components should not be interpreted as three chronological phases in artifact design. They are rather three 'aspects' of the overall transition from desired behavior to physical structure that are taken care of concurrently. I distinguish them for purely analytical purposes.

at hand? How to combine the structural possibilities for each of an artifact's desired behavior into a physically feasible whole that is affordable, efficient, convenient, etc. to boot?

The in-principle underdetermination is greatly aggravated by the fact that initial descriptions of desired behavior tend to imprecise, incomplete, and formulated at fairly high levels of abstraction.[13] For instance, 'a device that allows a person to alert someone else inside a house' or 'something with which a computer can be operated.' Such descriptions have to be made much more precise and the behavior(s) described in them must be translated into the more concrete descriptions of physics and chemistry before designers can come up with a categorical description of a structure that has the desired disposition(s). The more incomplete, imprecise, and abstract a description of behavior, the larger the 'amount' of underdetermination; an incomplete, imprecise, and abstract description applies to many more things than a complete, precise, and concrete one. To deal with this designers have to turn the initial description of desired behavior into something complete, precise, and concrete. That is component (b) of the overall transition from desired behavior to physical structure.

The third component is the transition from an intentional to a non-intentional description. Of the three, this one is the least interesting from the current perspective. To see why, we have to realize what goes on with intentionality during the design process so that the intentional presuppositions inherent in the starting point have disappeared in the description of physical structure at the end. This is not particularly mysterious. It is the all too familiar process of moving from an intention to do or make something to whatever actually comes of this intention, or, more specific, from intending to create or find something that meets certain specifications to what you actually come up with. I intend to make a really good impression on my boss but end up making a fool of myself. I intend to paint a truthful copy of Rembrandt's *Night Watch* but the result looks more like the doodlings of a five-year old. You, on the other hand, intend to write the authoritative exegesis of Heidegger's philosophy of technology and actually pull it off. In exactly the same way, designers intend an artifact to be able to do so-and-so and then come up with an object that can actually do so-and-so – or sometimes not. Either way, nothing of philosophical mystery happens in this process that does not happen continuously in our everyday lives of planning and acting. Although the blueprint that results from

[13] Surely not all design assignments will be formulated in abstract terms. If the initial description of desired behavior is already very concrete, then this simplifies the design process somewhat. Design processes that do start with a highly abstract description of desired behavior form the hardest case, so henceforth I will focus on these them.

a design process considered in isolation is a purely physical description of an artifact, it is also a blueprint *of an artifact with specific intended behavior(s)*. Just as in the above examples, the intended effect or result of what is created sticks to the end product. As I also noted in the first chapter, the intentional context is not completely lost during design.

So what happens with regard to intentionality during design processes is basically that a designer acts upon her justified beliefs and so confronts them with reality. Under the innocuous assumptions that the designer is competent and benign, i.e., tends to have beliefs with positive epistemic status about the components and materials she is working with and their dispositions and interactions, and does not expressly want to mess up design tasks, she devises a description of an object that she believes to have the right dispositions to perform the intended behavior(s). (In doing so, she has to grapple with the underdetermination problems mentioned above.) If she is correct, the intended disposition(s) will be among the many actual dispositions of the artifact and it will be able to behave as intended. If some of her beliefs turn out to be false, however, the intended disposition(s) will most likely not be among the many actual dispositions of the artifact described, in which case it will not be able to perform its intended function. The history of philosophy leaves no doubt that there are hard questions to ask about when beliefs have positive epistemic status and how they relate to reality, but answering those questions is the job of epistemologists and philosophers of science. Engineering design does not add anything new to the perennial philosophical conundrums of knowledge, belief, and truth.

So in the two chapters that follow I concentrate on transitions (a) and (b) of the overall design transition. How do designers turn an abstract description of behavior into a concrete description of the physical structure of an artifact in categorical terms? For the sake of brevity I will often use the expression 'behavior-structure transition' but henceforth this should be understood in the precise sense just outlined.

What happens in the context of explanation is in many ways the mirror image of design. Where design moves from a desired function to a physical structure, explanation uses the description of an artifact's physical structure to explain that an artifact has a function and can fulfill it. Therefore the three pairs of connotations I distinguished above are also relevant in the context of explanation. To account for an artifact's technical function on the basis of (among other things) its physical structure is to show how an intentional, abstract description of an object's behavior is true of a particular concrete physical object with so-and-so categorical properties.

Using the assumptions about technical functions introduced in section 2.3

we can immediately make a couple of important points about the explanatory connections between a technical function and an artifact's physical structure. First of all, accounting for an artifact's technical function cannot be done solely by means of its physical structure. A technical function is relative to a designer's beliefs and the social context in which the artifact's use plan is embedded. A function theory, such as the ICE-theory, makes this explicit. Second, if all is well and the artifact does not malfunction, an artifact's technical function is one of the things it can do; one of its possible behaviors (dispositions). Hence, to explain how it can fulfill its function, one needs to explain how an artifact can show certain behaviors or, equivalently, has the right dispositions. An explanation of this fact lies in the artifact's physical make-up. After all, objects can do what they can do in virtue of what they are made of, i.e., their properties categorically described.

The connection between physical structure and technical function thus splits in two. One part is the explanatory connection between physical structure and behavior (disposition) and the other part the connection between an artifact's many actual dispositions, its intended disposition(s), and its technical function. A function theory unearths this latter connection by analyzing a technical function as the disposition for which the artifact has been created or co-opted into a new use by a competent and benign designer or ditto users (and of which the function-ascriber believes (with positive epistemic status) certain things and which is embedded in a socially accepted use plan). I hesitate to call this connection explanatory since what a function theory provides is not an explanation by the lights of any available account of explanation. Giving the necessary and sufficient conditions for when function ascriptions are justified or true is an *analysis* of what it takes to have a technical function, rather than an *explanation* of a technical function.

As to the former connection: Just as any object's dispositions are explained by the object's physical structure, so too for artifacts. Such explanations go into the artifact's components and their exact physical make-up and properties. They explain an artifact's dispositions, which are sometimes described abstractly, by means of a very concrete and ultimately categorical description of that artifact's components and their properties. Since we saw that an artifact's technical function is the disposition for which it was created or selected, this type of explanation can serve as explanations of how an artifact can fulfill its function. They then get drawn into the intentional perspective, so to speak.

We need to distinguish between two cases: well-functioning and malfunctioning artifacts. A well-functioning artifact actually has the disposition for which it was made or selected. The designer's intentions were successful and her beliefs about the artifact correctly represent the artifact's actual structure and possible behaviors. Hence, the disposition that is the

artifact's function can be explained by the artifact's actual physical structure and the behaviors that structure makes possible.[14] A malfunctioning artifact has a function but does not have the disposition associated with that function. So there is no explanation of how it has that disposition. *Ex hypothesi*, however, the artifact still has a particular technical function. We can still explain how the artifact's designer (or some community of users) *intended* the artifact to have that disposition, i.e., how she thought the disposition would be realized by the particular physical structure she came up with. This explanation will look a lot like the one explaining an actual disposition of an artifact, except that it will be couched entirely in terms of what the designer (users) intended the artifact's dispositions and categorical properties to be. At least one of the beliefs about the artifact that are mentioned in such an explanation has to be false or at least incomplete.

I argue these points in more detail in chapters 6 and 7, where I develop a general account of how an artifact's behaviors can be explained (chapter 6) and then connect that account with a function theory to show how such explanations can become explanations of how an artifact can fulfill its technical function (chapter 7). To close the present discussion of my second main question, I want to get back to the three pairs of connotations identified above. In the context of explanation, just as in design, the two contrasts between abstract-concrete and dispositional-categorical are bridged at the same time. This happens in the type of explanations I just outlined. A concrete and categorical description of the artifact's physical structure explains an abstractly described disposition. A function theory takes care of the intentional-physical contrast by connecting a designer's (or users') intentions and beliefs with the physical structure and actual dispositions of the artifact. The intentionality pinpointed by a function theory also carries over to explanations. Explanations of the disposition that is an artifact's function are essentially explanations of an *intended* disposition, but this fact does not affect their general structure in significant ways.

## 2.5 Dispositional and Categorical Descriptions of Artifacts

The previous section introduced the distinction between dispositional and categorical descriptions of artifacts; descriptions that are about an artifact's

[14] Strictly speaking, the explanation recounts a designer's (or someone else's) *beliefs* about the artifact physical structure and its behaviors. If we assume these beliefs to be true, they form a transparent window on the artifact's actual structure and behaviors. But since our beliefs are fallible, we are sometimes mistaken about why an artifact has the dispositions it has. In that case, we end up with a false explanation of how an artifact can fulfill its function. Since this possibility is inherent to our epistemic condition, I will further ignore it here.

behavior and those that are not about behavior. This section serves to present a precise account of that distinction.

Anyone who is but faintly familiar with the philosophical debates about dispositions will know that the dispositional-categorical distinction is controversial. That should not prevent us from using it however, since pretty much the same holds for any distinction of philosophical import. Instead of agonizing over the controversies – let alone trying to sort them out – I will adopt one particular construal and use that to analyze my own philosophical questions. This move makes my work conditional on the correctness of that proposal, at least to the extent that what I say depends directly on it. This may seem an undesirable consequence, but I think 'inevitable' is more accurate. This thesis is not about dispositions, so I must help myself to some set of basic assumptions that I cannot argue for at length. As long as these assumptions are not in plain contradiction with common sense and engineering practice, or philosophically highly outlandish, any set of assumptions should be as good as the next. I admit that I do have my reasons for picking out Stephen Mumford's (1998) account of dispositions rather than one of the other options (such as Ryle 1990; Armstrong 1968; Prior *et al.* 1982; Prior 1985; Lewis 1997), but I wish to point out that these reasons do not need to be particularly strong in order for my procedure to be legitimate. As long as Mumford's view is considered to be a respectable philosophical option, as it certainly is, I feel free to use it in my project. I do not mean to be priggish about this point, however, so let me briefly give my reasons for what they are worth. (1) As far as I can see Mumford presents a compelling case for his views and (2) his account allows me to develop a plausible and fruitful interpretation of what designing engineers say about what happens in the transition from desired behavior to physical structure. As to the first reason, I must refer the reader to Mumford's own work and as to the second, I take chapters 4 and 6 to offer proof of that.

Let us now go over the basic elements of Mumford's account of the dispositional-categorical distinction. I will give more details on a need to know basis in chapters 4 and 6. His view has three components: an account of the dispositional-categorical distinction, a monistic property ontology, and a functionalist understanding of dispositions. We take them in turn.

What distinguishes disposition ascriptions from categorical ascriptions is their conceptual connection with subjunctive conditionals (i.e., claims about what an object *would* do, *were* certain conditions to be fulfilled). Disposition ascriptions are connected with subjunctive conditionals about an object's behavior *as a matter of conceptual necessity*,[15] while categorical ascriptions have

[15] This strong connection is required to protect the distinction from D.H. Mellor's objection that categorical ascriptions entail conditionals in the exact same way as dispositional ascriptions (Mellor

this connection only contingently, *a posteriori* (Mumford 1998: 77). To ascribe a disposition such as fragility to an object commits one to a conditional of the sort: 'Were the object to be dropped, it would break.' To ascribe a categorical property to an object, say triangularity, commits one to no such conditional. As it stands, however, this is too crude. Presumably a lot of things that we would not normally be inclined to call fragile would nonetheless break if dropped sufficiently hard. But the point of a disposition ascription is not to say that there exist bizarre circumstances in which the object will indeed show the behavior ascribed to it. It is rather to say that, under conditions relevant to the disposition ascription, the object can be expected to show the behavior ascribed to it. If it fails to do so, we want a special explanation for that. Mumford calls these conditions *ideal conditions* and takes them to be determined by the context of the disposition ascription (ibid.: 88ff.). A disposition ascription (in the actual world) is relative to the conditions of the actual world and to what can vaguely be referred to as 'normal' conditions. So the simple toy conditional above must be modified to include ideal conditions: 'Under context-determined ideal conditions, were the object to be dropped, it would break.'

The second element of Mumford's account is his property monism. As we just saw, Mumford defends a sharp distinction between dispositional and categorical ascriptions. He thinks this distinction resides at the level of our concepts and descriptions of the world. It does not carry over to the level of ontology as a distinction between two different kinds of properties. Properties come in just one kind, but we can pick them out by means of dispositional or categorical descriptions: 'There are not distinct categorical and dispositional types of properties but there are distinct categorical and dispositional ways of talking about instantiated properties' (ibid.: 145). Mumford supports this view with a token-identity theory on which any instance of a disposition is identical with an instance of one of its categorical bases. The solubility of a lump of sugar is identical with the molecular structure of that lump of sugar. His form of property monism is non-eliminativist and non-reductionist. Once it is understood that the dispositional and the categorical are really only two styles of denoting the same properties, there is no need to eliminate dispositional properties in favor of categorical ones (nor vice versa) and there is no need to reduce one to the other (ibid.: 172–191).

The third and final element of Mumford's account is its functionalist character. Mumford draws on resources collected by functionalism in the

1974, 1982). Mumford's reply is that, although it is true that categorical ascriptions are also connected with conditionals, they are so only contingently. The connection depends on the contingencies of our world and can only be known *a posteriori*. Not so for dispositional ascription, which necessarily entail conditionals (for details, see chapter 4 of this thesis and §4.6 of Mumford 1998).

philosophy of mind to articulate his precise understanding of dispositions. It is worth quoting him at length to see what this disposition functionalism amounts to:

> (i) What it is that makes a property or state *d* of an object *x* a dispositional property or state is that it is a conceptual truth that *d* causally mediates from stimulus events (for which *x* is patient) to manifestation events (for which *x* is agent).
> (ii) What it is that makes a disposition *d* the type of disposition it is is the specific stimulus and manifestation events to which it bears the relation of causal mediation. (ibid.: 197–198)

Clause (i) stipulates what Mumford means by functional: it is to mediate between stimulus and response events, between an input and an output. It is, in other words, to fulfill a designated causal role. This use is inspired by how 'functional' is used in functionalism in the philosophy of mind (see, e.g., Putnam 1975; Lycan 1999; Sober 1999). 'Functional,' as designating a description in terms of stimulus and response, is not directly connected to the richer notion of technical function, which, as we saw, presupposes intentionality. A functional description of an artifact in Mumford's sense is not a description of its technical function, but one that describes the artifact – more accurate: one or more of its properties – as mediating between a stimulus and manifestation event.

Accompanying this sense of 'functional' is a minimally committal and slightly idiosyncratic notion of function. To avoid confusion, call those functions 'Mumford-functions' or 'M-functions' for short. Property instances, or states, have M-functions, says Mumford (ibid.: 209–210), *relative to the predicate term used to denote them*. Relative to the predicate 'solubility' the M-function of the state denoted by 'solubility' is to mediate between immersion in water and dissolving. But relative to the predicate 'sweet', which denotes the same state, the M-function of that state is to taste a certain way on the tongue. That is because it is in virtue of tasting sweet on the tongue that we denote that state with 'sweet' and not in virtue of its mediating between being immersed and dissolving. Let me try to rephrase this. Calling sugar sweet bestows it with an M-function; the M-function to taste a certain way on the tongue. It bestows *that* M-function instead of, say, mediating between being immersed and dissolving because it is in virtue of sugar's tasting sweet on the tongue that we call it sweet and not in virtue of its dissolving in water. Referring to any object in dispositional terms bestows it with an M-function, namely the M-function to mediate between stimulus and response events – not just any stimulus and

response events, but those conceptually entailed by the dispositional ascription in question.

As I said, this notion of M-function is idiosyncratic. For two reasons: (1) We do not normally call dissolving in water the function of sugar; in fact, we do not talk about non-biological natural things as having functions at all. (2) The notion is much more liberal than notions of function in biology, social science, and technology, which apply only under more stringent conditions.[16] Most birds' beaks have the disposition to clack when rapped against the bark of a tree. Ascribing this disposition would suffice to confer an M-function upon beaks. But of course a biological function requires more; for instance an evolutionary account of how ancestors were selected for making clacking noises with their beaks or how clacking presently contributes to survival. For most birds and beaks, there is no such story to be told and therefore clacking is not a biological function of beaks (woodpeckers could be an exception, but I am not enough of an ornithologist to know for sure). It is hard to object to stipulative definitions, but it should be clear that the the notion of an M-function is not quite the same as the more robust notions of function that philosophers of biology, social science, and technology worry about.

This completes the overview of Mumford's account of dispositions. To prevent terminological confusion I will continue to use 'M-function' to designate Mumford's notion of function. Since 'functional' might be thought to equivocate between 'mediating between stimulus and response' and 'related to a technical function' I will try to avoid the term altogether and instead use 'dispositional' or 'behavioral' to refer to descriptions that are about causal roles and behavior. Exceptions will be explicitly indicated.

Perhaps an assiduous reader wonders whether my proposal to analyze one aspect of the function-structure contrast as the distinction between the dispositional and the categorical sits well with Mumford's functionalist account of dispositions. It might seem that there is some sort of circularity lurking in this conjunction of views. Fortunately, this is not the case. To explain why I need to make a few subtle points that take up more space than would be good for this chapter, which is already long enough. Therefore I address this worry in the appendix to this chapter.

## 2.6 Concluding Remarks

I have presented more precise accounts of design processes and of technical functions in this chapter. That allowed us to clarify the two central questions

[16] It looks as if Mumford is not fully aware of this point, since he introduces his notion by analogy with biological and technical functions.

further. To close this chapter, here are succinct and precise new versions of those questions:

(1) How does the design process realize a transition from an abstract dispositional description of an artifact's desired behavior to a concrete and categorical description of an artifact, i.e., a blueprint of an artifact's structure?
(2a) What do explanations that explain an artifact's abstractly described disposition(s) on the basis of its concrete physical structure look like?
(2b) How can the explanations from (2a) be combined with a function theory, so as to connect an artifact's physical structure with its technical function?

Question (1) is addressed in the two chapters that follow. My answer to question (2a) is given in chapter 6, with some preliminary work being done in chapter 5. Chapter 7 deals with (2b).

**Appendix: A Worry About Circularity Dismissed**

This appendix serves to address a worry about circularity I raised at the end of section 2.5. In the words of a possible objector, the worry could go like this. 'You have started out by saying that artifacts have technical functions and that they belong in an intentional perspective. This is shown, you say, by the fact that artifact functions presuppose intentionality and are subject to normative evaluations. In this chapter you said that the contrast between technical function and physical structure in the contexts of design and explanation of artifacts in fact has three aspects: (a) dispositional-categorical, (b) abstract-concrete, and (c) intentional-physical. Fair enough. But then, since the dispositional-categorical distinction is not so clear either, you adopt a specific account of it: Mumford's functionalist theory of dispositions. If you ask me, that lands you in a conceptual circle. You go from function to disposition to function.'

This is an eminently fair question. On the face of it, it might indeed look like I have constructed a clear case of circular reasoning. In actual fact, of course, I think this is not so. To see that, we can start with the following observation. For my analysis to be viciously circular, it would have to be the case that *analysandum* notion of function is the same as the *analysans* notion of function. In his comments on an article by Pieter Vermaas (2006) who also analyzes technical functions in terms of dispositions (which Vermaas calls capacities), Mumford himself says it well:

> The problem with this view [i.e., the view that analyzes technical functions in terms of capacities, JdR] is that one would then require an account of what a capacity was and this would have to be an account that did not use *the same notion of function* [my italics], upon danger of circularity.' (Mumford 2006: 79)

The key question, then, is whether the analyzed notion of technical function is the same as the analyzing notion of M-function. The answer to this question, I submit, is negative.

Earlier in the chapter I outlined my assumptions about technical functions and briefly discussed the ICE-theory of technical functions. This theory basically says that an artifact's technical function is the disposition for which the artifact was created or selected. This disposition is embedded in a socially accepted use plan. Anyone who ascribes this disposition as a function to the artifact must be justified in believing certain things about how the artifact's physical structure and the intended disposition are related and how the disposition contributes to the goal of the use plan.

We also saw what M-functions are. Something has an M-function, or a property of something has an M-function, when it is being referred to in functional terms, which is to say in terms of a causal role it can fulfill. For instance, dissolving in water is the M-function of sugar when we say sugar is soluble because 'soluble' is connected by conceptual necessity with the causal role of mediating between being immersed and dissolving in water. When we refer to sugar as being sweet, however, tasting sweet on the tongue is its M-function, because it is in virtue of *that* causal role, rather than that of mediating between being immersed and dissolving or any other, that we call sugar sweet. It is as simple as that; whenever we start talking about what things can do, their causal roles, M-functions are everywhere around us.

To those who worry that M-functions introduce or presuppose occult teleological powers in our otherwise neatly naturalistic worldview, I can say that that is not so. The only teleology M-functions require – that which makes one particular causal role rather than another an object's M-function – is entirely in our heads, as it should be on a naturalistic understanding of the world. When we denote some property in dispositional terms, we single out the causal role that that property occupies as its M-function, but that does not at all imply that that property has this function by itself, independent of our describing it. Only when we describe a property or object does it become possible to speak of its M-function.

It has to be admitted that this functionalist account of dispositions amounts to some reconstruction of our ordinary talk about dispositions and functions. We do not normally speak about dissolving in water or tasting sweet as *functions*

– in whatever sense – of sugar. Let alone of the bonding with two hydrogen atoms as a function of an oxygen atom. The only context in which such locutions might be imaginable is when sugar or oxygen atoms are intentionally used to accomplish some sort of goal, but even then it seems strained to insist on those uses as somehow really being their functions. In my humble opinion Mumford would have done well to be a bit more forthcoming about this reconstructionist edge of his view. He introduces his M-functions by analogy with biological and social functions, but does not explicitly state that his M-functions extend the domain of application of notions of function far beyond their normal boundaries to non-living natural things and stuff.

It should also be said, however, that there are other considerations that do recommend Mumford's view. There may be a general notion of function that Mumford tries to capture, which boils down to the notion of intended or actual causal role. If indeed there is such a notion, then technical functions, biological functions, and social functions are simply restricted subclasses of it. Mumford would then be right in making his analogy and extending function talk to natural non-living matter. However this may be, our goal is not to evaluate the merits of a functionalist understanding of dispositions, we only need to know whether such an understanding is compatible with the analysis of technical functions as dispositions.

We have now collected all the information needed to see that my proposal is not circular. I divide up my reply in two points. The first is that, as I said in section 2.4, I focus on artifact design as opposed to the encompassing use plan design. Artifact design starts with intended behavior. If the designer is competent and benign, and the design process is successfully brought to an end, this intended behavior becomes the artifact's technical function and the artifact can actually fulfill its function, i.e., actually has the intended disposition. The dispositional description with which artifact design starts, then, is simply a description of this desired technical function. The point of artifact design is to come up with a structural description of an artifact that will actually have this disposition or M-function.

I think it is obvious that there is no circularity involved in what I just said. Artifact design starts with desired behavior, i.e., a desired disposition, i.e., a desired causal role, i.e., a desired M-function. Its end product is a description of an artifact that has a great many dispositions, which entails it has a great many possible causal roles or M-functions. If the artifact can fulfill its technical function, one or a few of these actual dispositions stand apart as its technical function(s) in the sense specified by a function theory such as the ICE-theory. If the artifact cannot fulfill its function even though the designer was competent and benign and the design process was competently carried out, the intended behavior or dispositions with which the design process began is its technical

function, also in the sense specified by the ICE-theory or similar theory.

Even if there is no circularity involved in this dynamic description of artifact design, someone may still suspect there is a conceptual circle involved in the conjunction of the ICE-theory, my clarifications of the technical function vs. physical structure contrast, and Mumford's functionalist theory of dispositions. I do not think so and that is my second point. The ICE-theory or a similar function theory shows that an artifact's technical function is the behavior/disposition for which it was made or selected. Plugging in Mumford's account of dispositions, this means that a function theory shows that an artifact's function is the property for which it was made or selected under a dispositional description. And this amounts to it showing that a technical function is an intended M-function of an artifact. (As I noted repeatedly above, if the artifact can fulfill its function, it actually has this intended M-function, along with many other M-functions. If it cannot perform its function, it still has many other M-functions, but not this intended one.) Hence, the analysis goes from technical function to a dispositional description of properties to M-function. Is that circular? No, because technical functions and M-functions are, as we have seen, different notions. The former applies under far more restrictive conditions than the latter and thus has a much narrower domain of application. In fact, we saw reason to think technical functions are a subset of M-functions because M-functions are just causal roles. If that is correct, it is hardly surprising that an analysis of technical functions should end in M-functions, for there is no denying that a technical function is some sort of (intended) causal role.

CHAPTER III

# Design Methodology: What It Does and Does not Do

## 3.1 Introduction

The strategic assertions from the first chapter to the effect that the philosophy of technology would do well to take engineering practice more seriously must, of course, be fleshed out by actual study of engineering work. The present chapter is an attempt to make some headway with this strategy. I will look at design methodology to see what it has to offer for philosophers who are interested in engineering design in general and for philosophers such as myself who want to understand specifically how the transition from a description of desired behavior to physical structure is made. The rationale for this focus on design methodology lies in design methodology's self-ascribed goals of describing what design processes are all about and prescribing how they can be carried out more effectively and efficiently. I assume that if one wants to learn more about the essentials of design, the work of design methodologists is a natural place to start. It will turn out that this working assumption is tenable, but only in a qualified sense. Design methodology is a natural and perhaps even necessary point of departure, but it cannot be the final destination for a philosopher who is asking questions about design. Why this is so will become clear in what follows.

I should start by fending off one type of radical criticism that people have sometimes confronted me with when I told them that I turned to design methodology to learn about design processes. I deal with it here because this criticism would, if sound, undermine the entire project of this chapter. Some people deem design methodology completely irrelevant for design practice; they think 'real' design has little or nothing to learn from the neatly structured phase models, systematic methods, and streamlined techniques of design methodology. I do not share this extremely skeptical attitude. It may be true – though surely not always – that actual design processes are a lot messier than design methodology books have it, but that does not automatically invalidate the models, methods, and techniques of design methodology. There may be all sorts of reasons for why actual design practice is more muddled than can be glanced from the pages of design methodology handbooks. To name but a few: The models are not meant to describe actual practice, but to reform it; designers could be particularly bad at following sound advice; the advice may be rather

difficult to follow so attempts to do so invariably look messier than the advice itself; the models abstract away from irrelevant details; actual practice combines several methods and techniques at the same time. The inadequacy of design methodologists' models, methods, and techniques is but one of those possible reasons and it is hardly credible that it would be the only one or even the dominant one. I, for one, find it hard to believe that an entire discipline could sustain itself by producing irrelevant results. My attitude will therefore be one of prudent trust. I assume that a lot of the precepts of design methodology are genuinely justified, useful, and indeed conducive to successful design. I will add some qualifications to this later in the chapter.

With these fences up, let me present an overview of what is coming. This chapter has three major parts. In the first part, I give a brief historical overview of the field of design methodology (section 3.2). In the second part I zoom in on the work of three influential design methodologists, to wit Herbert Simon, Donald Schön, and John Gero, to see if they address the behavior-structure transition directly (sections 3.3 through 3.5). Since this turns out to be a dead end, I take a step back and look for more indirect ways to learn about the function-structure transition from design methodology in the third part (section 3.6). That task requires some additional reflections on philosophical methodology and brings along various observations about the field of design methodology from a philosopher's perspective. In section 3.7 I briefly get back to Simon, Schön, and Gero to see if my general observations about design methodology apply to them too. In the final section, I take stock and prepare the way for the next chapter.

## 3.2 A Brief Historical Survey of Design Methodology[17]

Design methodology is a young field, only about five-odd decades old. Nigel Cross (1984) distinguishes four historical stages or characteristic types of research in its development. It was not until the second half of the 20th century that designers started seriously pondering questions about how to systematize and thereby improve design. Part of their reason for doing so was a feeling of discontent with traditional design methods that relied chiefly on intuition and experience. Increasing technological complexity and faster change led to simultaneous increasing complexity and changes in the design task. Inspired by new and exciting insights in methods of problem solving, management, and operations research, which had come out of research during and after World War II, pioneering design methodologists set out to improve design processes. Their work constituted the first stage or research strand in design

[17] The historical sketch in this section relies heavily on Nigel Cross's (1984) helpful overview of developments in design methodology.

methodology. The common concern was to make the design process more efficient and reliable; less prone to error, re-design, and delay. Their avenue into improvement was to subject design to conceptual, logical, and mathematical analysis; to 'rationalize' it and make it more systematic. The immediate results of their efforts consisted of analyses of design processes into their (supposed) phases, identifications of crucial tasks that have to be carried out in the course of design processes, descriptions of various sorts of information that must be collected and processed in order for designs to be successful, and various tools, techniques, and methods to facilitate these phases and tasks, and to aid the required information gathering and processing. Representative work from this period was collected in the proceedings of the first conference on design methods (Jones and Thornley 1963). Some early design methodology books include Jones's *Design Methods: Seeds of Human Futures* (Jones 1970) and Asimow's *Introduction to Design* (Asimow 1962).

It soon became obvious, however, that the pioneering work, although seminal and stimulating in many respects, had been overly optimistic and perhaps even slightly naïve. Design processes and problems turned out to be much more complex than had been thought. Simple phase models did not make much sense in the face of often messy real-world design processes. A few key tasks did not capture the cascade of difficult issues and questions that designers had to confront in actual design. And a better-informed design process turned out to be not necessarily more efficient or effective. Such realizations led subsequent design methodologists to exchange the prescriptive stance of the early pioneers for a more modest descriptive one. This shift in focus led to a second and third research strand; the second focusing on the design *task* and the third on the design *activity*. The former explored the complexity and 'ill-structuredness' of design problems and the latter investigated how designers actually grapple with design problems.

Important contributions that fit squarely in the second research strand are Simon's work on ill-structured problems (Simon 1973) and Rittel and Webber's (1984) idea of a wicked problem. Their suggestion is that design problems are particularly hard because they lack structure in a number of important respects. Among the characteristics of wicked problems that Rittel and Webber mention are the following:

(a) A definite formulation of the problem does not exist, formulation is part of the problem;
(b) There is no stopping rule, it is unclear when a solution has been reached;
(c) Solutions are not true or false, but good or bad;
(d) There is no immediate and ultimate test of a solution;

(e) The set of potential solutions is not enumerable or exhaustively describable;
(f) Every problem is essentially unique.
(Rittel and Webber 1984: 136–144)

Of course, their claim is not that each and every design problem will be a wicked problem, but some design tasks certainly are. Simon adds that there is no clear-cut distinction between ill-structured or wicked problems and well-structured or tame problems; they are extremes on a continuum, allowing for infinitely many shades of structuredness and wickedness. He also suggests that ill-structured problems are iteratively transformed into well-structured ones by adding additional criteria, requirements, and constraints and decomposing the problem into more manageable, i.e., better-structured, sub-problems (Simon 1973). Although this type of research abandoned the earlier prescriptive stance, hope remained that better understanding of design problems would lay the foundations for tools, methods, and techniques to improve design.

The same goes for the people who worked, and are still working, in the third strand that emerged from the early work on design methodology. They, too, aimed to improve design in the long run. To accomplish their goal they focused on what designers actually do when they are designing. Design was seen as one particular type of human behavior and studied accordingly with the methods of empirical psychology and cognitive science. Later on, when the preponderance and importance of team work in designing was recognized, sociological and anthropological research methods were employed to scrutinize design activity. In the course of this research, vast amounts of empirical data were gathered, ranging from interviews with designers, protocol analyses of designers thinking out loud, to questionnaires, reports and statistical data on designers' solutions and strategies in more controlled experimental or laboratory settings with standardized design tasks. The role of drawings, sketches, prototypes, etc. has been investigated; there have been studies of design dialogues with clients and co-designers, the influence of design teams and forms of cooperation, groupthink, conflict, creativity methods, and so on and so forth. It is well nigh impossible to summarize the results of this ever-growing research tradition, but if anything has been firmly established, then it must be that real design processes are a good deal messier than the systematic models of design processes and activities have it. As Cross aptly observes, however, this is not necessarily a criticism, because systematic models were never meant to mimic actual practice, but instead to transform it.

When these two descriptive projects had been underway for some time, a more reflective and fundamental approach to design methodology arose: the fourth stage or research strand. People started reflecting on the fundamental

concepts in design and cautiously tried to take some steps towards a renewed and more informed prescriptive stance in design methodology. Whilst these people recognized the importance of paying close attention to findings that came out of descriptive empirical work, they also saw that mere observation and description is not easily going to lead to recommendations for improving design. What is needed in addition, they proposed, is an adequate theory of design, which can undergird systematic procedures and techniques to improve design. Research of this ilk is therefore no longer directly concerned with designers, design activity, or methods and tools for design, but with the basic concepts relevant for design, such as design knowledge, experience, cognitive schemata, rationality, abduction, etc. Two major contributions to this tradition, with which we will make closer acquaintance later in this chapter, are Herbert Simon's *The Sciences of the Artificial* (1996) and Donald Schön's *The Reflective Practitioner* (1983). Both books outline an encompassing view – some even say a paradigm – of what design is. These views do not directly address questions about how to design more efficiently, effectively, or reliably but have inspired many others to develop answers to such questions. My own work in this and the next chapter is closely related to this last research tradition in design methodology. Although I suspect many design methodologists may find it too 'philosophical', in the sense of it being highly abstract and theoretical and thereby a little too far removed from actual design, its goal is clarification of some of the basic notions that more practically minded design methodologists regularly and, it seems to me, somewhat unreflectively employ. I will say more about the exact relation between design methodology and my project later in the chapter.

Now I want to add a few remarks to this brief historical sketch. The four stages I described are not mere historical periods. They also shaped substantively the field design methodology by introducing characteristic styles and topics of research in design methodology that are still alive and kicking today. Hence, design methodology contains (1) prescriptive research aimed at outlining ideal design processes, (2) descriptive research that investigate design tasks and problems, (3) descriptive research that observes design processes and designer behavior, and (4) reflective or foundational research that develops, clarifies, and explicates the fundamental concepts of design.

That brings me to the next point. The historical overview does not yet tell us what design methodology is and what its characteristic goals are. In fact, an answer to those questions is harder to find than one may expect, because most design methodologists – presumably not unlike other scientists – hardly bother to reflect on the nature of their field and its characteristic goals, approaches, and problems. Instead, they simply *do* design methodology by launching straight into the projects that make up design methodology as a discipline. Exceptions

to this rule are few, but not impossible to find. In line with the historical developments in design methodology, Nigel Cross proposes the following characterization:

> [Design methodology] is the study of the principles, practices and procedures of design in a rather broad and general sense. Its central concern is with how designing both *is* and *might be* conducted. This concern therefore includes the study of how designers work and think; the establishment of appropriate structures for the design process; the development and application of new design methods, techniques, and procedures; and reflection on the nature and extent of design knowledge and its application to design problems. (Cross 1984: vii–viii)

This characterization is in agreement with another more recent one by the Dutch design methodologists Norbert Roozenburg and Johan Eekels. Their characterization is more explicit in distinguishing a normative or prescriptive goal of design methodology on the one hand and a descriptive goal on the other. They also add that the overall goal of design methodology should be improvement of the effectiveness and efficiency of design process.

> In design methodology there are two principal questions: (a) what is the essential structure of designing? and (b) how should the design process be approached to make it effective and efficient? It is the task of descriptive design methodology to answer the first question, and the second should be answered by prescriptive design methodology. Descriptive design methodology tries to reveal the methods applied in design through logical structural analyses, and empirical research, as well as to identify the needs for methodical support. Prescriptive or normative design methodology forms an opinion based on descriptive analyses, and recommends for certain problems the application of certain methods, or even demands it. (Roozenburg and Eekels 1995: 29–30)

One can find roughly similar thoughts between the lines of the introductions and forewords to many textbooks on design methodology. Design methodology, then, seeks both to describe what design is and how it is done, and to prescribe what it should be and how it should be done. To accomplish these aims, it employs all research methods it sees fit.

Before I move to the next section, I want to mention one further development in design methodology that has added a new strand to the existing research in recent decades. It is research that revolves around artificial intelligence and design. The idea of designing robots and computers appeals to many.

As a result, researchers have tried to computerize various design tasks. The aims range from highly ambitious attempts to let a computer replace human designers through computers doing well-structured sub-tasks to computers figuring as fancy design aids to human designers that can perform complex information processing tasks and calculations.

With all this we have a fairly decent insight in what design methodology is all about. That means we can now go back to the first main question for this thesis – how designers get from an abstract description of desired behavior to a physical structure in design processes – to see if design methodology provides answers to that question. In the following three sections I look at three influential theoretically-minded design methodologists and investigate whether their work contains resources to elucidate the behavior-structure transition.

### 3.3 Herbert Simon: Design as Rational Problem Solving

Since the question I am asking is philosophical and theoretical, it stands to reason that we should be looking at design methodologists who are working in the theoretical-fundamental fourth research strand. Although this limits the possibilities somewhat, there are still plenty of design methodologists that could be relevant. My choice to look at Herbert Simon, Donald Schön, and John Gero is therefore to some extent arbitrary and reflects the contingencies of my personal research history. The following additional bit of history, however, does provide some justification for the selection. We pick up where we left the last section, i.e., with the fourth historical stage that introduced more fundamental-theoretical approaches to design. According to Kees Dorst (1997) this tradition has painted two fundamentally different pictures of design: Herbert Simon portrays design as a species of rational problem solving and Donald Schön depicts it as a specific type of reflective practice. Dorst even speaks of two paradigms in Thomas Kuhn's sense. One may quibble about the aptness of this nomenclature, but it will be convenient to have a term to refer to these different frameworks so I will follow his usage. Because the two paradigms are so different they span a spectrum of possible positions and, in so doing, may well be considered as representative for a large part of theoretical design methodology.

Simon's and Schön's work dates back to the 1970s and 1980s, so one might think although *they* did not worry about the sort of questions I am asking, some of their successors did. This is not the case, however. I will also look at a more contemporary figure, John Gero. He is also relevant because of his keen interest in artificial intelligence. His work therefore gives us a peek into that sector of design methodology. We will look at Simon in this section, at Schön in the next, and at Gero in the one after that.

Herbert Simon's reflections on design comprise ideas about design problems, the designer, and the design process. The primary source for his ideas is of course his seminal *The Sciences of the Artificial* (1996). We already encountered his work about design problems as ill-structured problems in section 3.2, so there is no need to repeat that here. Simon models designers as goal-seeking information processors who are subject to certain cognitive limitations. In the face of the complexity of design problems and as a result of their cognitive limitations, designers will typically *satisfice* rather than optimize, i.e., settle for a solution that can be discovered through a practically feasible search and that is good enough, as opposed to optimal (Simon 1996: 119). The design process is a specific instance of a human rational problem solving process. Problem solving, says Simon, is 'a search through a vast maze of possibilities, a maze that describes the environment. Successful problem solving involves searching the maze selectively and reducing it to manageable proportions' (ibid.: 54). For design, in particular, the relevant problem is to devise an artifact that fulfills particular practical goals. This characterization can be made more specific by inserting Simon's characterization of properly functioning artifacts:

> An artifact can be thought of as a meeting point – an 'interface' in today's terms – between an 'inner' environment, the substance and organization of the artifact itself, and an 'outer' environment, the surroundings in which it operates. If the inner environment is appropriate to the outer environment, or vice versa, the artifact will serve its intended purpose. (ibid.: 6)

A designing engineer's task is thus to come up with a structure, or inner environment, that has all the right capacities to fulfill its intended function properly in the circumstances in which it is to be used, i.e., is well-adapted to its outer environment.

How are designers supposed to attain this adaptation? The very short answer is: optimizing and satisficing. According to Simon, a design problem must be translated into a mathematical optimization problem, which can then be solved by means of an appropriate search algorithm. The inner environment is represented by a set of actions or things the designer has control over. Sometimes this set is specified in its entirety, but usually it is given by a number of command variables ranging over some domain. These variables stand for the characteristics of the artifact the designer can determine. The outer environment enters the design problem as a set of parameters, either with fixed value or as possible values within some interval, either deterministic or following some probability distribution. Since the goal is adaptation, the next thing required is a mathematical function that measures the successfulness of the adaptation. Of course the command variables and the environment

parameters go into this function, but it is usually supplemented with additional constraints that exclude certain combinations of command variables and environment parameters. The result is a design optimization problem: Find admissible values for the command variables that optimize the adaptation function under the given environment parameters and constraints. All available mathematical resources pertaining to optimization can be brought to bear on design problems. This is also where satisficing comes in. If an optimization problem is intractable, as many design problems likely are, mathematically less demanding satisficing techniques may be able to find non-optimal but still satisfactory solutions.

This approach of Simon's presupposes that the relevant design alternatives are readily given. That supposition is clearly at odds with many practical design situations, where designers have to go through great lengths to find new alternatives. No account of design can be complete without addressing this issue. Enter means-ends analysis. In the course of their life people acquire vast amounts of information about how actions affect the world, or, in other words, how actions can be means to ends. Provided one has a particular goal in mind a cleverly organized search through this stock of experiential knowledge will often yield various actions or sequences of actions, new or tried, that will produce the desired state of affairs when carried out successfully and in the right manner. This is what means-ends analysis does: a clever search through pairs of means-ends to discover new design alternatives, which can subsequently be translated into the language of the optimization problem as command variables.

To reduce the complexity of design tasks, Simon recommends decomposition of the problem into relatively independent sub-problems. The behavior of a complex artifact can usually be broken down into relatively independent sub-behaviors. The design of the components that take care of these independent sub-behaviors can then be carried out with some degree of independence of one another.

That completes Simon's account of the design process. Before we move on to see if it speaks to my philosophical question, it will be helpful to summarize it by rephrasing its central points in my behavior-structure terminology (for a similar translation of Simon's work see Kroes 2002). Design, Simon agrees, is about finding the appropriate structure for an artifact that will then be able to fulfill a specific desired function, i.e., contribute to the realization of some practical goal in relevant circumstances. Finding this structure is essentially an optimization process, the point of which is to find a structure that optimizes the 'fit' between the artifact and its circumstances, i.e., which behaves optimally in the given circumstances and subject to the specified constraints. If it is necessary to come up with new solutions, designers should dig into their

practical experience to find possible structures (means) that will lead to the desired goals.

What are we to think of this account? Simon's work has been hugely influential and spawned a large body of literature both inside and outside the field of design methodology. One of the first things that come to mind is to what extent his account matches up with data gathered in empirical design studies. The evidence weighs pretty heavily against Simon here. Designers use formal optimization methods to solve well-structured and tractable sub-problems, but many design processes involve much more than that, especially those in which innovative artifacts are developed. In his *What Engineers Know and How They Know It* (1990), aerospace engineer Walter Vincenti gives his readers an in-depth look at the multifarious activities that make up actual design processes. A lot of what goes on in them has to do with hard-to-formalize, 'soft', qualitative, and intuitive types of knowledge. It is quite a stretch to believe that, deep down inside, all of that is or could be reduced to formal optimization and satisficing methods. I think we should say, with the benefit of hindsight, that Simon got his hopes up too high regarding the potential applicability of utility theory and optimizing methods to real-world design problems.[18]

Empirical adequacy, however, is not the all-important issue for me. Even if it were true that all design is optimizing and satisficing, I would still want to know how the transition from a dispositional description to a categorical one is accomplished in the course of this process. To that question we must now turn. Simon does not address it directly, but his theory of design does contain hints from which we can try to reconstruct an answer. If anywhere, the transition must happen in the process of finding the right sort of inner environment or in finding the right means to a desired end. The reason why it does not surface is, I believe, that it is lost in translation. Simon's 'inner environment' and 'means' equivocate between a categorical and a dispositional reading. The inner environment, or 'substance and organization of the artifact itself,' can be described in terms of behavior, black-boxing the details of the implementation, or in categorical terms, specifying materials, spatial properties, etc. The same goes for means. A time-traveling machine would make an excellent means to the end of time travel, but this dispositional description of a machine tells you absolutely nothing about possible structures of actual time-traveling machines. This is not to say that means are always described in dispositional terms; they may just as well be described in categorical terms. My point is that the mere technique of means-ends analysis does not automatically take care of the

[18] Hampton (1994) forcefully pushes the objection that utility theory is deeply flawed as an actual or ideal model for human rationality. If that objection holds water, Simon's project would be thoroughly undermined. That discussion, however, is wide outside the scope of this thesis.

transition from behavior to structure. Rather, that transition seems to have been smuggled in through a backdoor and is implicitly presupposed by Simon.

As a last resort, one may hope that, at the same time designers master the technique of means-ends analysis, they also pick up the abilities to make the behavior-structure transition so that the way in which they do the former also makes it plain how they do the latter. According to Simon, means-ends analysis is based on experiential knowledge about what consequences ensue when people perform actions. So how do designers learn means-ends analysis? The answer boils down to: 'By experience.' That sounds right, but in a disconcertingly one-size-fits-all way. It is rather similar to someone who, when asked how to do philosophy, says: 'By thinking very hard.' So this, too, is a dead end if we want to learn something more about the transition from behavior to structure.

Simon's work thus has little to say about the transition from behavior to structure; it just presupposes that this transition happens somewhere in a design process. It does contain a range of precepts about tools and techniques designers ought to use. Further investigation may reveal if and how these enable the behavior-structure transition; I will say more about this possibility after we are done discussing Schön and Gero.

### 3.4 Donald Schön: Design as Reflective Practice

If Herbert Simon represents the rationalist, systematic branch of design methodology, Donald Schön stands for those design methodologists who emphasize intuition, implicit knowledge, creativity, and other 'softer' aspects of design. Like Simon, Schön's interests were broader than just design; he wanted to know how highly skilled professionals put their expertise to use in actual practice. His seminal book *The Reflective Practitioner* (1983) attacks the standard-paradigm of technical rationality – also Simon's view – according to which professional activity 'consists in instrumental problem solving made rigorous by the application of scientific theory and technique' (ibid.: 21). Schön rejects this paradigm for various sorts of reasons and seeks to replace it with that of the *reflective practitioner*. Key to that paradigm are the notions of *knowledge-in-action*, 'a kind of knowing [that is] inherent in intelligent action' (ibid.: 50), and *reflection-in-action*, '[thinking] about doing something while doing it' (ibid.: 54). Philosophers will be quick to recognize undertones of Gilbert Ryle's know-how (1990) and Michael Polanyi's (1958) tacit knowledge. Schön indeed credits them for inspiration.

Knowledge-in-action is a kind of implicit know-how that is expressed in and through the actions of professionals. Professionals know how to carry out certain actions, recognize salient aspects, or pass judgments without having to think about them explicitly prior to carrying them out. They are typically

unaware of their knowledge-in-action. Knowledge-in-action is hard to put into words, let alone formalized propositions or methods; expressions such as 'getting a feel for' or 'finding the groove' are what professionals come up with to describe their knowing-in-action. Reflection-in-action is elicited when professionals start thinking about their knowing-in-action and do so while performing these actions, often initiated by an unexpected happening or consequence. At such moments, practitioners may start to form hypotheses about what accounts for their success or failure and why they do what they do; they form theories-in-action. Reflection-in-action also serves to correct less successful internalized strategies for action.

It would be a mistake to pit Schön against Simon as if the latter is the champion of systematicity and rationality while the former advocates free-flowing anarchism and irrational intuition. Instead, Schön urges that a serious look at professional practitioners reveals that their rationality breaks out of the narrow confines of instrumentalism and utility theory. It is this broader and empirically more adequate conception of rationality that he tries to capture. He does not champion irrationality but a different, broader, notion of rationality.

An important goal of Schön's paradigm is to provide an antidote to the prevailing role of the technical rationality paradigm in engineering curricula. Reflection-in-action ought to play a huge role in education, Schön submits (see especially Schön 1987). The rigorous science and formal methods one typically finds dominating engineering courses are all very well, but they do not endow students with the required internalized knowledge-in-action. Schön's program for education, in contrast, has experienced professionals/teachers make explicit their reflection-in-action on their own and their students' work. By doing so they convey knowledge-in-action and train students to become reflective practitioners as well.

Schön applies his paradigm to design situations in the third chapter of *The Reflective Practitioner* and in later papers (Schön 1984, 1988, 1992; Schön and Wiggins 1992). I will concentrate on the basics of his account of design. Design, Schön says repeatedly, is a 'reflective conversation with the materials of a design situation.' Unlike Simon, he makes no particular assumptions about the designer. He finds fault in Simon's explication of design tasks, which presupposes too much pre-given structure. Simon's science of design only works for well-formed problems that have been abstracted away from actual practice. Schön takes more of a constructivist stance towards design tasks. Every design task is one of a kind, a 'universe of one,' and design tasks have no inherent structure of themselves. Designers impose structure through acts of *naming* and *framing*, i.e., highlighting salient features of design situations and seeing them as having a particular structure, thereby opening up avenues of possible solutions. There is no telling in advance if an act of framing will work

out, but expert designers' knowledge-in-action will cause them to be fairly reliable at framing design problems in fruitful ways. The next phase is for designers to make *moves*, which means to perform thought experiments with potential (partial) solutions to see if they lead to a satisfactory solution that fits the situation. Moves are continually *evaluated* in the light of the design problem and so is the fit between the chosen framing and the design problem. Sometimes thought experiments lead to unexpected new consequences and insights. Schön calls this *back-talk* of the situation. Skilled designers appreciate and listen to back-talk and use it to perform new thought experiments or to *rename* and *reframe* the design task or aspects of it and start a new cycle of naming-framing-moving-evaluating. The cycles come to an end when the designer is satisfied with her solution.

The reception of Schön's account has been a mixed bag. On the one hand, many designers have felt that Schön's description of design processes strikes just the right balance between mere description and reconstruction. They can identify strongly with Schön's talk of naming, framing, moving, back-talk, and evaluation, as well as with the general reflective practitioner paradigm. It evokes the right connotations and fits their own lived design experience. On the other hand, the majority of design methodologists have greeted Schön's account rather coolly. Although they appreciate his attempt to describe design in a fundamentally different way, there is widespread consensus that his account is too vague and ambiguous to put to any practical use. Dorst summarizes the situation: 'Schön developed a 'primer' for a new theory of design, and as such the basic theory of reflective practice is rather sketchy; the key concepts are vague and its uses are not totally clear' (Dorst 1997: 73).

With that observation, we have also identified the key difficulty that arises when we try to figure out what Schön's account tells us about the transition from behavior to structure. At first, one may even wonder to what sorts of design processes the account is supposed to apply, for it may seem to be more appropriate for design processes in which the softer, aesthetic considerations take center stage. Perhaps Schön's account captures architectural and industrial design quite well, but not, say, mechanical engineering, electrical engineering, or aerospace engineering. That is not what Schön had in mind, however. He intends his account to be fully general, describing the 'fundamental process underlying differences among design professions' (Schön 1983: 77). There is serious reason to call this generality into question, but let us not worry about that too much. We will assume for the sake of argument that Schön's ideas also apply to the early phases of design in the more technical engineering disciplines.

What, then, can they teach us about how designers make the transition from behavior to structure? That is hard to tell. The transition must happen

somewhere in the reflective conversation, but apart from iterations of naming-framing-moving-evaluating this design conversation has little structure of its own that could indicate at what point(s) the transition takes place. Furthermore, acts of naming and framing involve both imposing structure on the design problem and choosing to try out particular categories of possible solutions, which suggests that desired behavior and ideas about physical structure co-evolve during reflective conversations. Once we try to precisify Schön's vernacular in order to translate it into the terminology of behavior and structure, we get stuck immediately in its vagueness and ambiguity. For instance, how could we render the phenomenon of back-talk intelligible? Schön might mean that a designer suddenly discovers that the structure she had in mind does not show the intended behavior as well as she expected. Or that the intended behavior turns out to be impossible to realize in any feasible structure. Or that the designer discovers a new detail about the intended behavior of which she was previously unaware and which opens up the way to different possible structures. Or that she notices how a particular behavior or sub-behavior is actually similar to one with which she is familiar. The thing is, we have no idea which of these he means and, what is more, we have no way of settling the issue. Pretty much the same holds for Schön's entire account. It may be rich in meaning and highly evocative but attempts to make it more precise run afoul of its vagueness and ambiguity. Besides, it also looks as if on any of the proposed interpretations either Schön presupposes that designers are able to make the transition from behavior to structure in the right way or the phenomenon he describes does not clearly relate to that transition. Either way, we do not make any progress in answering philosophical questions about the behavior-structure gap. To actually make this case would of course require more careful exegesis of Schön's writings, but I trust that the argument sketch I have provided suffices to make my conclusion plausible.

The conclusion of this section is rather similar to that of the previous one. Like Simon, Schön has little to contribute to a philosophical account of the transition from behavior to structure. He appears to presuppose that that transition is made somewhere in a design process without explicating how this is done. The chief difficulty with his work lies in its vagueness and ambiguity. The fact that his suggestive but elusive wordings appeal to so many designers does strongly suggest that design processes contain an ineliminable core of creativity and intuition that might be very hard if not impossible to put into words. While I emphatically take notice of this point, I will not pursue it further, since I am interested in those parts of the design process that can be explicated through philosophical analysis.

### 3.5 John Gero: Function, Behavior, Structure

The third and final design methodologist we will be looking at is John Gero. Over the last two decades Gero and his coworkers[19] at the Sydney-based *Key Centre of Design Computing and Cognition*[20] have been immensely prolific in developing and refining a basic model of engineering design and putting that model to work in numerous computer applications that aim to simulate, assist, and improve aspects of design. Apart from the fact that he has a well-developed model of design, Gero is especially interesting from my perspective because his model appears to be a perfect match for my project. He frames design in exactly the same way as I do, as a transition from function or desired behavior to structure.[21] Over the years, his model has been modified and extended repeatedly, but most of the alterations have preserved a common core. Tracing all of these would take up much more space than I have here, so I will focus on the core of the model and briefly look at the few fundamental revisions Gero has proposed.

Gero first introduced his Function-Behavior-Structure model (henceforth FBS model) in a 1990 paper in *AI Magazine*. That paper is still a good place to start an exposition of the FBS model.[22] Design, says Gero, is an activity in which a set of desired functions *F* is transformed into a description *D* of one or more artifacts that can perform these functions. Clients and their needs provide *F*. *D* must be sufficiently detailed for engineers to be able to manufacture the artifact(s) described in it. The fundamental problem in design is that there is no direct route from function to structure (or vice versa, for that matter). Any function allows for different structures and any structure can perform many different functions. There is mutual underdetermination. 'It is well recognized that there is no function in structure and that there is no structure in function,' as Gero observes (1990: 31).

Functions are transformed into design descriptions through a series of elementary steps. Gero distinguishes five basic steps, which are supplemented with three loop-back steps allowing for iterations (for the latter, see Gero and Kannengieser 2002). The first step is *formulation*. It translates the function(s) *F* into expected behavior, $B_e$, for an artifact that will perform *F*. Second is *synthesis*,

[19] For the sake of brevity, I will henceforth use only Gero's name to refer to this collective.

[20] For more information see its website: *www.arch.usyd.edu.au/kcdc*.

[21] We will see later that Gero's notion of function, and thereby the exact nature of this transition in his model, has changed quite a bit over time. We do not violate his intentions, however, if for the time being we take his words to point at the transition from an abstract description of desired behavior to a concrete and categorical description of an artifact.

[22] Besides Gero's own work, I also draw on the helpful summaries of his ideas provided by Pieter Vermaas en Kees Dorst (Dorst and Vermaas 2005; Vermaas and Dorst 2006).

***Table 1*** *Design Steps in the* FBS *model*

| | | | |
|---|---|---|---|
| Step 1 | Formulation | $F \rightarrow B_e$ | Translation of desired function(s) into behaviors that realize the function(s) |
| Step 2 | Synthesis | $B_e \rightarrow S$ | Transformation of expected behaviors into prospective structure of artifact |
| Step 3 | Analysis | $S \rightarrow B_S$ | Derivation of actual behaviors of proposed structure |
| Step 4 | Evaluation | $B_e \rightarrow B_S$ | Comparison of expected and actual behaviors |
| Step 5 | Documentation | $S \rightarrow D$ | Production of design description |
| Step 6 | Reformulation 1 | $S \rightarrow S'$ | Change of proposed structure |
| Step 7 | Reformulation 2 | $S \rightarrow B_e'$ | Change of expected behaviors |
| Step 8 | Reformulation 3 | $S \rightarrow F'$ | Change of desired function(s) |

which transforms $B_e$ into a prospective structure $S$ of the artifact. The third step is *analysis*: the actual behavior $B_S$ of an artifact with structure $S$ is derived. The fourth step is called *evaluation* and consists of a comparison of the expected and actual behavior, $B_e$ and $B_S$. If the result of this comparison is satisfactory, the process moves to the fifth step of *documentation*, in which a description $D$ of the artifact's structure $S$ is produced. If the result of step four is unsatisfactory, the process loops back to one of the earlier steps. A different structure $S'$ is proposed, the expected behavior is altered into something different, $B_e'$, or the desired functions are revised, resulting in a new set $F'$. Table 1 from Vermaas and Dorst (2006) summarizes the process.

The FBS model thus introduces behavior as an intermediary concept between function and structure. Since a direct transition from function to structure is not possible, the notion of behavior functions as a bridge between them. By translating both function and structure into behavior, an indirect transition is made possible, or so Gero thinks.[23]

In order for this model to work the steps Gero prescribes must be feasible.

[23] If you accept my assumptions about technical functions this is an odd idea. If a function is the

They should at least be easier than the direct transition from function to structure. This is not immediately obvious but of course Gero thinks that this is indeed the case. The reason why differs for each step. For the analysis step, $S \rightarrow B_S$, physics provides the required knowledge; physical theories provide the means to derive the behaviors of a given structure. The synthesis step, $B_e \rightarrow S$, is the reverse of the analysis step. Gero suggests that it is accomplished through an abductive inference. Philosophers may be tempted to call it an inference to the best structure. There are no theories or rules available for doing the formulation step, $F \rightarrow B_e$. Designers have to take recourse to experiential knowledge of past and learned design. Since Gero says nothing further about the evaluation and documentation steps, I take it that he assumes those to be unproblematic.

Interestingly, however, Gero also introduces another device to get through the design process: *design prototypes*. A design prototype is 'a conceptual schema for representing a class of a generalized heterogeneous grouping of elements derived from alike design cases that provides the basis for the start and continuation of a design' (Gero 1990: 30–31).[24] It is the basic unit of designers' experiential knowledge and brings together knowledge from design situations about functions, structures, and behaviors. Design prototypes come in various degrees of generality; from very general, like 'toy' or 'engineering structure,' to highly specific, like a particular type of screw or a specific brand and model laser printer. A prototype for a particular artifact contains (descriptions of) its functions, its expected and actual behaviors, and its structure. It further includes (1) relational knowledge about associations between $F$, $B_e$, $B_S$, and $S$; (2) qualitative knowledge about the effects modifications in $S$ have on $F$ and $B$; (3) contextual knowledge relevant to the design situation such as values of exogenous variables; and (4) computational knowledge about mathematical relations between $S$, $B$, and possibly contextual variables.

According to Gero, prototypes help designers to make the formulation step ($F \rightarrow B_e$), the synthesis step ($B_e \rightarrow S$), the analysis step ($S \rightarrow B_S$), and the three reformulation steps ($S \rightarrow S'$, $S \rightarrow B_e'$, and $S \rightarrow F'$). Here is how. First, prototypes are retrieved based on the desired function(s) $F$. The set of prototypes so retrieved contains a wealth of information about functions, behaviors, and structures from which designers must then select the relevant bits. The second process is therefore selection of relevant prototypes. The information contained in the relevant remaining prototypes provides the epistemic means for carrying out the formulation and synthesis steps. In the course of these steps, the remaining

intended behavior of an artifact, behavior cannot play an intermediary role between function and structure. The intuitions behind Gero's model must be different than mine.

[24] 'Design prototype' is thus a term of art for Gero. It is not supposed to stand for scale models or test versions of an artifact as it does in its normal use.

prototypes are pruned; all the information that is irrelevant to the design task at hand is left out. Next, one or more prototypes are instantiated. That is the third process. Fourth and finally, the instantiated prototypes are refined. This happens through analysis and evaluation steps in which the actual behaviors of the instance(s) are derived and compared with the expected behaviors. If the results are unsatisfactory reformulations steps are carried out. All of this is again supported by the knowledge contained in the prototypes.

Gero's model provides an extensive and detailed account of design. Moreover, the model frames design problems in a way that is very similar or even identical to how I framed them. It would seem, then, that his model gives a direct answer to my philosophical question. Designers get from function or desired behavior to structure through the described steps and they use design prototypes to support these steps. For the sake of argument we will assume that the story about design prototypes does not conflict with the earlier description of how the design steps are taken.[25] I will take it that the earlier story is more general and that design prototypes are a specification of that story.

Unfortunately, things are not as easy as they seem. First, a closer look at Gero's use of the central notions of function, structure, and behavior spells trouble. The definitions of these terms have changed several times and that makes it hard to see which of the steps in the model pertain to the transition I am interested in, from desired behavior to physical structure. In consequence Gero's exact account of that transition is hard to pinpoint. Even if this can be resolved by fixing the definitions for good, a second problem arises once we try to bring design prototypes into the fold. Prototypes threaten to trivialize the problem of the transition from behavior to structure and in the end it looks as if Gero has little to say about the transition from behavior to structure. We take these problems in turn.

Starting from the original 1990 paper Gero and varying colleagues have made three major attempts to define the central notions of the FBS model. They can be found in Gero *et al*. (1992), Rosenman and Gero (1998), and Gero (2002) respectively. Although the definition of structure has not changed substantially, that of function and that of behavior have. Rosenman and Gero (1998) add an additional concept, purpose, to the model, which has dropped out again in Gero (2002). Vermaas and Dorst (Dorst and Vermaas 2005; Vermaas and Dorst 2006) carry out a careful exegesis of the definitions. I will summarize their

[25] This assumption is not obviously true. Take, for instance, the suggestion that the analysis step ($S \rightarrow B_S$) is carried out with the help of physical theories. In the design prototypes story, this step is supported by the information contained in prototypes. That may imply that analysis is simply looking up the behaviors known to be exhibited by the structures in the prototypes. The outcome may be identical, but the processes certainly are not.

main points here; details can be found in the original papers. They show that function swings back and forth between a largely intentional reading in which it is equated with purposes and intentions and a more physical reading in which it refers just to the physical dispositions and behaviors of artifacts. Behavior is a mixed notion; it has fluctuated from being a purely physical notion, to more of an intentional notion closely akin to technical function, and later back to a physical notion. By way of example, here are Gero's original definitions (Gero *et al.* 1992: 193), which are very similar to the most recent ones in Gero (2002):

> Function: the design intentions or purposes
> Behavior: how the structure of an artifact achieves its functions
> Structure: the components which make up an artifact and their relationships

These definitions are rather imprecise by philosophical standards, but that is not our chief concern here.[26] It is at least reasonably clear where the transition we are looking at is supposed to happen. If functions are purposes, then the formulation step ($F \rightarrow B_e$) becomes a transition from a goal to a desired behavior. That is to say, in my terms, it becomes use plan design. The transition I am interested in then happens primarily in the synthesis step ($B_e \rightarrow S$), and perhaps also in the analysis ($S \rightarrow B_S$) and evaluation ($B_e \rightarrow B_S$) steps. Designers make this transition through abductive inferences, 'inferences to the best structure.' They are aided in this by the knowledge collected in prototypes. But now look at the following revised definitions from Rosenman and Gero (1998):

> Function: the results of the artifact's behavior
> Behavior: the artifact's actions or processes in given circumstances of the natural environment
> Structure: the elements of the artifact, the material arrangement of these elements and their connectivity

Again, I think there are issues with interpretation and precision here, but let us leave those for what they are.[27] Function and behavior are now much closer to

[26] For instance, 'how the structure of an artifact achieves its functions' equivocates between a largely dispositional reading in terms of behaviors of the artifact and its components and a more categorical reading in which the properties and materials that make the behaviors possible are described. Furthermore, the definitions say nothing about how abstract or concrete functions and behaviors are to be described.

[27] Take the expression 'the results of an artifact's behavior.' A natural reading of that would be to interpret it as referring to some state of affairs, but that leads to the highly counterintuitive

one another and therefore the formulation step ($F \rightarrow B_e$) may also be a part of the transition we are studying. Gero said that there are no rules or theories available for carrying out this step; designers have to take recourse to experiential knowledge, which is presumably stored in prototypes. As I said, the latest definitions (in Gero 2002) put purposes back into functions, thus changing the formulation step back into being a kind of use plan design. So depending on the set of definitions one takes, the transition from desired behavior to a categorical description of physical structure happens in different places and by slightly different epistemic means.

Of course there is nothing wrong in principle with changing one's opinion, but there are reasons to suppose that the latest change in definitions is not going be the last one. Vermaas and Dorst (2006) list two. First, the latest definition of function gratuitously collapses the distinction between function and purpose (as did the original one); functions are defined as an artifact's purposes. But consider unsuccessful use of well-functioning artifacts. Our everyday ways of talking about this indicates that there are plenty of cases in which an artifact fails to achieve its purpose but still functions properly. Looking at the LCD-screen of a digital camcorder in bright sunlight is a case in point. Although the screen may function perfectly, its purpose is not achieved. This simple observation suggests that Gero's present definitions are still unsatisfactory. The second reason has to do with conceptual continuity across domains. Since Gero ties function to purposes, his notion of function is automatically discontinuous with the domain of biology, which employs, according to most, a non-teleological notion of function. This is not a knockdown objection, but the familiar philosophical dictum that one should not multiply entities beyond necessity dictates that it is preferable to explore the prospects for a non-domain-specific notion first. Vermaas and Dorst also offer a constructive proposal to that end, which is in line with my own preferred way of cashing out the distinctions between function, behavior, and structure.

The second problem is basically this: On any of the proposed readings of his definitions, what Gero says about the transition that interests us is rather too general to provide any accurate understanding. Taking all options together, the transition from desired behavior to a physical structure is made by abductive inferences and application of experiential knowledge, both supported by design prototypes. This is fine, but I will now argue that design prototypes cloud the view we may have hoped to get on the transition from behavior to structure. Design prototypes summarize and systematize a designer's experiential design knowledge by bringing together functions, behaviors, structures, and various

conclusion that an artifact's function is some state of affairs. Philosophers would of course never talk about an artifact's *actions* as actions presuppose intentionality.

epistemic relations between them. They forge inductive connections between functions, behaviors, and structures. Some combinations of function and structure or behavior and structure will be regularly conjoined in design. Such conjunctions then license inductive inferences from new desired behaviors to structures and behaviors that have been known to be associated with those behaviors.

Note how this line of reasoning has smuggled in connections between behavior and structure through the general backdoor of past experience. Gero more or less explicitly admits that this is indeed what design prototypes do. To the bold assertion that there is 'no function in structure and no structure in function,' he adds that 'human and design experience produces a connection between function and structure' (1990: 31). But then the very issue that Gero started out with becomes moot, in a way not unlike we saw happen in Simon's work. Experiential knowledge represented in design prototypes fashions the required connection between behavior and structure and thereby trivializes the transition from one to the other in one fell swoop. It may be perfectly true that experience relates intended behavior and physical structure, but I would hope that there are more specific things to say about how this happens than what Simon and Gero offer.

In closing this lengthy section, let me lay out its upshot. In spite of its good looks, Gero's FBS model does not deliver the philosophical goods. Its central notions, function, behavior, and structure, do not (yet) have stable and satisfactory definitions, making it hard to pinpoint what Gero has to say about the behavior-structure transition. With the introduction of design prototypes, we even lose sight of our transition, because they effectively short-circuit it. In the end, then, the FBS model no more enlightens the philosophical problem of how designers are able to make the transition from desired behavior to physical structure than did Simon's and Schön's accounts of design.

To be sure, this is a bit of a disappointment. What to do next? One possibility is to continue looking for other design methodologists in the hope of eventually finding one who does address my philosophical questions. That will not be my path of choice because I am somewhat pessimistic about its prospects. If we accept Kees Dorst's claim that Simon and Schön represent two paradigms in design methodology, we have a reason to think their insensitivity to the behavior-structure gap are representative for a large part of design methodology. And to the extent that Gero is representative for another bunch of design methodologists, perhaps the artificial intelligence enthusiasts, they too will have little to offer by way of answers to my question. In spite of my pessimism, though, I would be happy to hear of design methodologists who do deal with the transition from behavior to structure to a satisfactory degree of specificity. Another possibility, the one I will pursue below, is to look at the

practical pieces of advice design methodologists offer. If this advice is sound – as I will assume it is under conditions to be specified in the next section – it should help designers to make the transition from behavior to structure. And if that is so, a closer look at what this advice says must reveal something about how designers are able to make the transition. The next section elaborates this strategy and deals with some of the problems it faces. Chapter 4 then puts it to work.

### 3.6 Design Methodology and the Philosophy of Technology

The direct approach of the previous three sections fails. The design methodologists I looked at had nothing specific to say about how designers get from behavior to structure. Therefore I propose we take a step back to reconsider the role design methodology can play in my philosophical project. To carve out this role, I will make a few general evaluative observations concerning design methodology.

Before I offer these observations, however, a caveat about their status. What I have to say runs the risk of being misunderstood as being unfair or even plain confused when it is interpreted as criticism of design methodology. That would be uncharitable, however, since I am not trying to criticize anyone. Rather, I will be pointing out that what design methodology has to offer does not suffice when one asks philosophical questions about design. It provides valuable input for philosophical theorizing, but does not offer a philosophically satisfactory account of the transition from behavior to structure. This should hardly come as a surprise, however, since the standards for success in philosophy and design methodology are very different. Good design methodology was never meant to deal with philosophical issues. What I argue below, then, is no more a criticism of design methodology than it would be a criticism of physics to point out that it does not answer theological questions.

I want to expand on this point a bit further because it is important to fully appreciate it before proceeding. Like design methodologists, I am concerned with the question of how designing engineers make the transition from function to structure. Unlike design methodologists, however, I address this question from an analytic-philosophical perspective. Negatively, this means that I am not trying to give a fully accurate description of all the empirical facts of design, nor am I trying to devise and prescribe new steps, methods, tools, or techniques that will speed up and improve design. Positively, what I want is an account of how designers can make the transition from behavior to structure, i.e., how their actions and thinking pave a path through the underdetermination of physical structure by abstractly described intended behavior. Since the more theoretical design methodologists from the previous sections did not clarify this process, my alternative plan is to see if the practical

prescripts of design methodology, or at least some of them, embody successful strategies to deal with this underdetermination. To put the same point in other words, design methodology says what designers do and should do, but I am interested in why what design methodology says is supposed to work; what the rationale is for thinking that the methods, tools, and techniques of design methodology will contribute to solutions of design problems. For an analogy, think of the algorithms for doing divisions with large numbers that kids learn in primary school. It is one thing to develop these algorithms, describe them, and to instruct kids in how to use them properly, but quite another thing to show that they indeed realize a correct way of solving division problems. Design methodologists are chiefly concerned with the former sort of tasks but I want to take on the latter.

The upshot is that what I say below is not an attempt to identify shortcomings on the part of design methodologists – even though some of my remarks will have a critical flavor. I do not presume to be a design methodology expert. For all I know, there may be perfectly good reasons for why design methodologist do what they do. Without exception, my remarks are driven by *philosophical* interests. They concern things that strike me, as a philosopher looking at design methodology from the outside, as obscurities, ambiguities, or omissions. At the same time, they clarify how my philosophical project interacts with the deliverances of design methodology; what it must incorporate, where it is left to its own devices, and what sort of departures from design methodology are legitimate or even necessary.

My new strategy – to analyze practical precepts to see how they embody successful strategies to get from behavior to structure – gives rise to two questions: (1) Which precepts should be studied and (2) to what extent may philosophical analysis reinterpret and reconstruct what design methodologists say? I split up the first question in two sub-issues: (a) Which precepts are well supported by the empirical data of actual design and (b) how can we distinguish normative precepts from descriptive claims, given that design methodologists do not always draw a clear distinction?

I start with the empirical support issue. A lot of design methodology handbooks virtually ignore the empirical support for their precepts. More often than not, one finds but a few hand-waving remarks in the foreword or introduction about 'empirical research and years of practical experience' to support what the rest of the book has to offer (see, for instance, Asimow 1962; Jones 1970; Hubka and Eder 1988; Ullman 1992; Roozenburg and Eekels 1995; Pahl and Beitz 1996). This by itself does not imply that empirical support does not exist and that such books deal in extensive speculation about design processes, but it does indicate that design methodologists do not make a huge

point of systematically discussing the relation between empirical, descriptive studies of design and practical experience on the one hand and the claims of prescriptive or normative design methodology on the other.

This omission would, perhaps, be forgivable as long as design methodologists are in universal agreement. In that case one might be justified to conclude that the empirical evidence, whatever it is, at least points in the same direction. Unfortunately, however, this is not how things are. There is plenty of debate among design methodologists, both about the niceties of various techniques and about larger issues such as the very usefulness and feasibility of particular methods, the required tasks in a design process, and the information required for them. In fact, at a more fundamental level some even contest the value of systematic design methods, emphasizing the importance of intuition and creativity. For a relative outsider to a field it is always hard to find out which positions are better backed up by the empirical evidence, but since design methodologists tend to be less than forthcoming about the empirical support for their ideas, deciding whom to believe becomes next to impossible.

So where does that leave me, as a student of design methodology? In choosing particular methods to study in more detail I am inevitably aligning myself with some design methodologists and discarding the views of others. This is not as bad as it may seem, for in spite of their many disagreements design methodologists do agree on some issues – albeit not universally. Some precepts are 'common denominators,' i.e., a version of them is recommended by most design methodologists. For instance, there is widespread agreement that a design process includes and should include phases like problem analysis, conceptual design, detailed design, and evaluation. As for specific techniques, many methodologists write about goal trees, means-ends analysis, functional decomposition, morphological analysis, quality function deployment, and the house of quality, etc. Because these common denominators are so widely recognized, it seems reasonably safe to assume that they are backed up by solid empirical evidence and indeed contribute to successful design. By analyzing only such prescripts one avoids the unattractive possibility of ending up analyzing unfounded or non-representative ideas of individual design methodologists. This seems to be a prudent strategy for a philosopher who tries to consult design methodology. In the next chapter I investigate one particular widely recommended technique, functional decomposition.

The next issue has to do with the distinction between the descriptive and prescriptive branches of design methodology. As we saw earlier, design methodology comprises both descriptive and prescriptive projects. This is fine, as long as it is clear which bits are supposed to be descriptive and which prescriptive. But that is not always the case. Design methodology handbooks in particular often appear to occupy a suspect middle ground between the

descriptive and the prescriptive.[28] They will both claim that the models, methods, and techniques presented accurately describe the way designers actually work and then go on to claim that this is also how they ought to work.[29] Some of them nearly explicitly commit a naturalistic fallacy by claiming that, because so-and-so is how design is done, it is also how it should be done. For instance, having described the steps of the basic design cycle, Roozenburg and Eekels state:

> This is the way design processes are structured. That establishment leads almost imperatively to the statement: effective design processes should be structured in this manner. The cycle, which is found descriptively, changes into a norm for effective designing. We can therefore also consider the basic design cycle as a *prescriptive* model for designing. (Roozenburg and Eekels 1995: 93)

An outright dismissal of this practice of blurring the line between is and ought would be naïve. It may be the only feasible course of action for a normative design methodologist. I would even argue that, to some extent, the practice is justifiable. One may take two routes to defend it: one inductive and the other theoretical. The inductive route goes like this. If there are designers who are demonstrably more successful than their peers when judged by appropriate standards, then that is a good reason to investigate their behaviors.[30] If it turns out that they exhibit special behaviors, use certain techniques, or divide up the design process in specific phases, then these behaviors, techniques, and phases are likely to be conducive to good design and are therefore to be recommended as good practice. We already encountered the theoretical route in the historical overview of section 3.2; the fourth stage or research strand in the history of design methodology gave rise to theoretical accounts of the nature of design.

[28] Vermaas and Dorst (Dorst and Vermaas 2005; Vermaas and Dorst 2006) also make this point about Gero's work.

[29] The interest in artificially intelligent designing computers complicates matters even further, for that introduces yet more options between purely descriptive and purely prescriptive design methodology. It can be an aim to have computers mimic the actual behavior and thought processes of human designers, some form of supposedly ideal human behavior, the actual outcomes of human design processes, or perhaps some supposedly ideal outcome.

[30] It is an excellent question what these appropriate standards ought to be. I can only hint at possible answers. One simple answer could be to ask the community of designers or clients their opinions about the quality of the design. Another option requires a theoretical account of good design, but this option may well boil down to the second way of justifying prescriptive design methodology, to which we are about to turn.

An adequate theory of design may provide grounds for prescriptive claims about design methods. The idea would be that by understanding the nature of an activity, you automatically grasp something of how this activity can be carried out successfully. As an example, suppose your theory of design tells you that design is, deep down, nothing but the systematic reshuffling of existing components and parts. This theory would then be able to support normative claims to the effect that, say, designers ought to study the history of design, memorize parts of product catalogues, or create large databases of products, components, and their characteristics. An adequate theory of design would be inspired by and true to more accurate descriptions of design, which is of course why my toy-theory fails as a serious candidate.

So the practice of inferring normative claims from descriptive ones may be justifiable after all. What bothers me, then, is not so much this practice itself, as design methodologists' seeming unawareness of its problems and their failure to even note the urgency of a justification.[31] Design methodologists hardly ever mention the fact that underlying their precepts is an unarticulated leap from is to ought – let alone discuss this leap systematically. No work of design methodology I have come across contains any attempts to clarify how description may undergird prescription. Nothing is said about successful track records of the methods prescribed, evaluative standards that are supposed to separate the good designs and processes from the bad ones, or how theories lend support to what is offered. The fact that design methodologists hardly discuss the empirical data for their claims, as we saw above, only adds to this problem.

What is a philosopher of engineering design to make of this? The wisest thing to do, I submit, is not to wait for design methodologists to provide the required justifications, but to start somewhere and focus again on common denominator normative precepts. The commonness of these precepts at least counts as some inductive evidence for their being justified and conducive to successful design. Careful philosophical analysis of what design methodologists say about such precepts may then reveal further theoretical justification. (Or it may not, in which case we would need to think again about either the data of design methodology or our analysis.) This will therefore be my approach in the next chapter. I focus on one common denominator precept, the technique of functional decomposition, and assume it is normatively justified. The starting point for the analysis will be what design methodologists have to say about functional decomposition and its use. But before I spoil the plot of the next

[31] My complaint here does not obviously apply to more theoretically-minded design methodologists working in the tradition of the fourth stage or research strand. One would expect that their work does offer the required sorts of justifications.

chapter we ought to move on to the second question that my strategy needs to confront.

That question – how far philosophical analyses should be allowed to move away from the empirical data – touches on a very general philosophical issue and I can only scratch its surface here. Because philosophical analysis attempts to explicate and clarify concepts that are unclear or ambiguous, it invariably involves some amount of reconstruction and interpretation of the empirical data. Without that, there would hardly be a point to the analysis, for it would not provide any information that the empirical data did not already provide. At the same time, excessive amounts of reconstruction and seemingly far-fetched interpretations will greatly diminish the plausibility of an analysis. General rules to steer clear of both the Scylla of uninformative quasi-analysis that merely repeats the empirical data and the Charybdis of all-too-fantastic reconstruction do not exist. Each case must be judged by its own merits. For example, on the standard analysis of knowledge as justified true belief, an utterance like 'I know he did it, but I cannot believe it,' would make no sense, as knowledge implies belief. There can be little doubt that such utterances are part of the standard use of 'know,' but most epistemologists refuse to take this as a knockdown argument against the standard analysis. They prefer to reinterpret the data by, say, denying that statements of this sort are to be taken at face value or just dismissing them as sloppy speech. For an example closer to the other end of the spectrum, think of Peter Van Inwagen's notorious argument to the effect that material objects like tables and chairs do not, strictly speaking, exist (Van Inwagen 1990). In spite of the many philosophical virtues he claims for his view, many people consider the radical reconstruction of much of our everyday talk about tables and chairs it brings along as too high a price to pay and therefore reject his view as crazy.

For me the problem boils down to how I ought to treat what design methodologists say about their concepts, methods, and techniques. As the previous paragraphs showed, a straightforward reading already reveals difficult issues with regard to empirical support and the distinction between descriptive and normative claims. So even the 'empirical data' itself is not unambiguous. Putting this difficulty aside for the moment, I will summarize how I propose to analyze design methodology without over-analyzing it. The basic rule will always be that my analysis must follow and be consistent with what design methodologists say under charitable interpretation. Since design methodologists are not philosophers and hence usually say little on the ontological or epistemological commitments of their views, I will feel free to extrapolate possible positions from what they do say and choose a plausible position that is fruitful for my philosophical project. When my analysis leads to revisions or reinterpretations of design methodological claims, I consider

myself bound to do two things: (1) give an argument for why these revisions and reinterpretations are necessary and (2) account for why design methodologist would not adopt my revised or reinterpreted way of speaking about whatever is at issue.[32] I am acutely aware that these open-ended principles fall far short of a watertight solution to this methodological problem, but they are all I can offer.

This completes the outline of my alternative strategy to incorporate the findings of design methodology into my work. Before the rubber meets the road in the next chapter, two other things are still on the agenda: a quick look back at Simon, Schön, and Gero to see if they too are sloppy about empirical support and the distinction between descriptive and normative claims and a concluding summary of this chapter.

### 3.7 Simon, Schön, and Gero Revisited

Another look at Simon, Schön, and Gero lends further support to my observations about design methodology from the previous section. Although their work is of a more theoretical nature than that of the design methodologists I had in mind in the last section, it turns out that they are also unclear about empirical support and the distinction between descriptive and prescriptive claims.

In spite of its more fundamental nature Simon's work does contain claims about what design is (or should be). With regard to the first problem, empirical support, Simon appears to have a slanted view of the sort of empirical support he needs. He refers chiefly to artificial intelligence research, but hardly to empirical studies of design. This sort of evidence may foster a satisfactory account of how computers tackle design problems, but to claim that it can also ground a general account of design is quite a stretch. The second problem, lack of clarity about descriptive or prescriptive goals, does not directly affect Simon. He is pretty explicit about the normative aims of his theory; it outlines what design ought to look like and it seeks to reform design education accordingly. He also has the theory to back up this project. The problem, as we just noted, is that this theory is rather poorly supported by empirical facts about design, even if an optimistic interpretation of the results of artificial intelligence research at the time when Simon was working on *The Sciences of the Artificial* may have rendered it somewhat plausible. It is therefore questionable whether one should accept the normative claims of Simon's rational problem-solving paradigm.

Schön seems to do better with regard to empirical support. His reflective

[32] The introduction of a rational reconstruction of design processes and a precise account of technical functions in the previous chapter are cases in point. Both interpret and reconstruct what designers and methodologists say without, I think, contradicting them or violating their intentions. The benefit of adopting these analyses is added clarity and precision.

practice paradigm is based on extensive studies of professionals in action and the application to design of it is also supported by analysis of a review that a master designer gives of a student's work in an educational setting. Furthermore, because he analyzes how *successful* professionals work, it is plausible that his initially descriptive account can be extrapolated to a prescriptive one. Nonetheless, the bold claim that Schön's account is fully general and gets at the essence of all design processes is overhasty. The evidence he adduces is mainly from architectural design and can therefore not support so general a claim. Second, the distinction between description and prescription is not always clear. It often looks as if Schön does both at the same time. The justification for that may run along the lines I hinted at (descriptive analysis of successful professionals leads to prescription), but Schön does not make this explicit. Furthermore, because Schön seems to be running his descriptive and prescriptive projects together, it is unclear what his prescriptions are. It is one thing to observe certain phases in reflective conversations, but quite another to tell designers, say, how they can recognize these phases and what they should do in them. Schön does not offer much concrete help here and the vagueness and ambiguity of his account make it difficult to distil such help from it.

Gero's work also suffers from the same problems. First, empirical support for Gero's account in the form of real studies of designers at work in real design processes looks meager. That is one of the morals of Dorst and Vermaas (2005). The empirical data that Gero and coworkers tried to analyze in terms of the FBS model does not bear out the conclusion that the model is indeed an adequate description of what goes on (or ought to go on) in design processes. Meanwhile, Gero and his group have produced a truly impressive amount of empirical work in which the FBS model was used successfully to automate parts of the design process. That would seem to lend some support the conclusion that the model is a promising way of mimicking design processes in computers or other intelligent agent systems. But that is a different and not directly related claim. This also brings me to the second problem, the distinction between descriptive and prescriptive goals of the model. Like Simon and Schön, Gero embraces both goals with his model. He wants it to capture the core of what design is and what it ought to be. Like most design methodologists, however, he does not provide any justification for the transformation of his model, which is initially presented as descriptive, into a prescriptive one. Vermaas and Dorst (2006) level a similar criticism against Gero, arguing that because of its double status the model is subject to two evaluative regimes, which are hard to reconcile smoothly. As a descriptive model, it should expose what actually goes on in design processes, and as a prescriptive model, it should help designers improve their designs and processes. It is unclear why anyone should suppose that these two regimes go hand in hand. I would add that there is a third strand

running through Gero's work: the goal to build designing computers. This goal complicates matters further because it is far from obvious how descriptions and prescriptions about human designers will relate to descriptions of successful designing computers (see also note 29 above). At one point, Gero frames his work as an attempt to make computers more intelligent by letting them handle their tasks more similar to the way humans do (Gero and Rosenman 1990: 66–68). If that is indeed the overarching aim, one would expect him to put much more effort in psychological and sociological studies of design before moving on to computer modeling. I cannot help but wonder if Gero's actual research perhaps turns this aim on its head: it starts from a model that apparently has some success in computer applications and then claims that human designers work or should work more like this model. At any rate, Gero runs together description, prescription, and computerization in an unhelpful way without justifying that practice in any way. Hence he provides another illustration of the tendency I ascribed to design methodologists in general.

## 3.8 Conclusion

To conclude the chapter I will briefly retrace the steps of its main argument and, in doing so, formulate a few general guidelines for how an empirically informed philosophy of technology can benefit from the work of design methodologists.

The goal of this chapter was to find out to what extent design methodology provides useful input for an empirically informed philosophy of technology and, in particular, whether it contains solutions to the philosophical question of how designers are able to get from behavior to structure in design processes. A brief overview of the history and goals of the discipline of design methodology showed that it studies design processes with the overarching aim of improving them. Design methodologists conduct all sorts of empirical studies of designers and design processes, build descriptive and prescriptive models of design processes, develop models, methods, and techniques that assist designers in their work, and some reflect on the nature and fundamentals of design. On the face of it, then, design methodology looks like a good place to find answers to philosophical questions about design.

A closer look at some influential design methodologists who have proposed theoretical accounts of design tempered the initial enthusiasm. Herbert Simon, Donald Schön, and John Gero all make important points about design, but in the end they do not address the difficulties involved in bridging the behavior-structure gap. Both Simon and Schön presuppose that designers are able to make the transition and, in spite of first appearances to the contrary, Gero's work shows a subtler version of the same presupposition. Simon and Gero do say something about how designers acquire that ability, but their answer – 'by experience' – is too coarse-grained to be of any real help.

So design methodology does not contain direct answers to my philosophical question. Section 3.6 outlined a more indirect strategy for using design methodology in my project. The basic idea of this strategy is to investigate the practical advice design methodologists give in order to see how this advice is supposed to contribute to a successful transition from behavior to structure.

A number of problematic issues cropped up in the development of this strategy. First, as an outsider to the field, one finds oneself in an epistemic quandary of having to decide whose practical advice to believe without anything much to go on. Design methodologists hardly say what empirical support they can adduce for their views. This would be all right as long as they all agree, but that is not the case. So one cannot simply pick up a design methodology anthology and distil a consensus view about, say, the essential steps of design processes. To deal with this difficulty, I will focus on common denominator precepts. While this is not a fail-safe approach – consensus does not imply truth – it is the best thing an outsider can do.

The second issue is that design methodologists exhibit a tendency to jump to prescriptive claims without bothering much about how such claims can be grounded in descriptive accounts of design or theoretical considerations. There are ways to deal with this justification issue but design methodologists have not been eager to explore them. The best thing an outsider looking for well-grounded precepts can do, I submit, is again to concentrate on common denominator precepts and assume that they are justified and hence effective. To the extent that widely held opinion among experts is indicative of justification, such precepts have the best odds of being effective.

The role design methodology can play in my project, then, is twofold: supplier and recipient. When treated with appropriate selectivity, its practical advices provide the *analysanda* for my project. If specific tools, methods, and techniques contribute to effective design, analysis should reveal how they do so on the basis of theoretical considerations about their nature and precise workings. The results of such analyses can be fed back into design methodology to check their validity. When design methodology embraces such analyses, its theoretical basis is strengthened. This latter process is largely beyond my reach; it is the job of interested design methodologists. I can only work on the former task. That is precisely what I will do in the next chapter where I analyze the method of functional decomposition in detail and argue that it indeed helps to bridge the gap between an abstract dispositional descriptions of artifacts and concrete categorical ones.

# Functional Decomposition

## 4.1 Introduction

Functional decomposition is a widely recommended design method. My goal in this chapter is to see if we can understand why this is so, i.e., to see if we can understand how functional decomposition contributes to the transition from abstractly described behavior to a concrete categorical description of an artifact. To accomplish this goal, we will spend quite some time specifying what functional decomposition actually is, because it turns out that design methodologists exhibit a somewhat ambivalent attitude towards it. Many textbooks emphasize the importance of functional decomposition, analysis, or modeling, but this emphasis is hardly reflected in efforts to analyze and develop the notion systematically. More often than not, the student of engineering design finds herself with not much more to go on than a few sketchy examples of intuitively plausible functional decompositions. To my mind, this is an undesirable situation. If functional decomposition is indeed an important method and conducive to successful design, then we should try to understand why and how this is so.

There is another reason for looking specifically at functional decomposition, as opposed to other methods. The end product of a process of functional decomposition – *a* functional decomposition – plays an important role in the context of artifact explanations, as I will argue in chapter 6. Whatever comes out of the present chapter will therefore be of instrumental value for an account of such explanations.

Finally, functional decomposition also plays a role in discussions about complex systems, mechanisms, functions, etc. in the philosophy of biology, philosophy of science, and philosophy of mind. There is no guarantee that an investigation into the meaning of functional decomposition in the context of engineering design will be illuminating for other fields as well, but neither are there *prima facie* reasons to think that it is not. I will not have the time to explore such transdisciplinary connections, but I imagine that would be a worthwhile follow-up project.

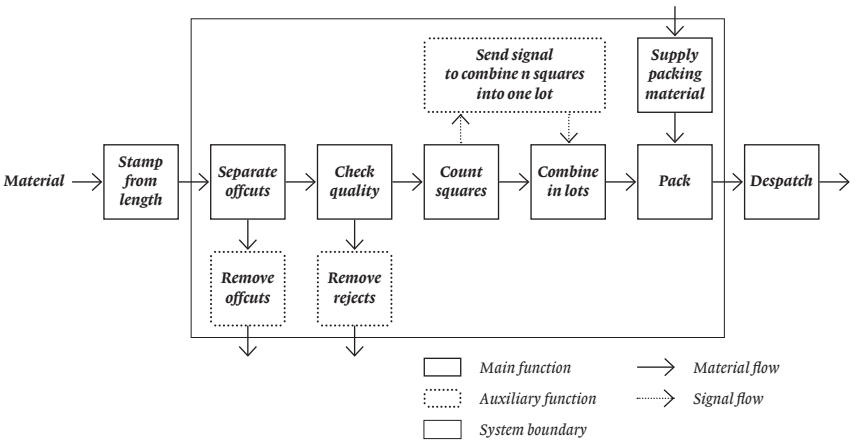
So much for the motivation behind what I will be doing presently. The rest of the chapter basically consists of a step-by-step analysis of functional decomposition, which ends up distinguishing a number of different senses

of functional decomposition. This analysis then allows me to explicate how functional decomposition bridges part of the gap between a dispositional description of an artifact and a categorical one. The outline for the chapter is as follows. I start with empirical data: examples of functional decompositions from design methodology textbooks in section 4.2. In section 4.3 I give an overview of some more sophisticated work related to functional decomposition. In the course of these two sections I identify several features of functional decomposition that need to be accounted for. In section 4.4 I apply the distinction between dispositional and categorical descriptions of artifacts to clarify and explain some of these features. Although this already provides valuable insights in how functional decomposition may fulfill its role in a design process, it does not yet lay bare the mechanism of functional decomposition. So in sections 4.5 through 4.8 I open up the black box of functional decomposition and argue that functional decomposition is really an umbrella term for a number of different processes that all contribute in their own way to the behavior-structure transition in design. In section 4.9 I put all the pieces together and outline my account of functional decomposition and show how this account underpins the role design methodologists ascribe to functional decomposition in design processes.

Terminological disclaimers: (1) In chapter 2 I promised to avoid the term 'functional' altogether because it might be thought to equivocate between 'mediating between stimulus and response' and 'related to a technical function' in the context of this thesis. This chapter makes it hard to keep that promise since the method we are investigating happens to be called '*functional* decomposition.' When design methodologists talk about functional decomposition and functions in the context of design, I interpret that as meaning decomposition and analysis of the intended behavior of the artifact being designed, i.e., of the intended behavior that is to become the artifact's technical function. (The examples below will show the legitimacy of my interpretation.) As I said in chapter 2, the reason why 'function' and 'functional' are often used without reservation in the context of design is probably that the distinction between (intended) behavior and function is unimportant in this context because an artifact's intended behavior normally becomes its function. Nonetheless, I will try to be as precise as I can. In summarizing and quoting the views of design methodologists in sections 4.2 and 4.3 I will occasionally have to follow their use of 'function' and 'functional' but when I am commenting on their views and throughout the rest of the chapter I will consistently use 'behavior' and 'dispositional' to refer to descriptions of an artifact in terms of its behaviors.

(2) 'Functional decomposition' stands for both the *process* of decomposing a given intended behavior and for the *result* of this process, i.e., a mental or other

***Figure 1*** *Function Structure for Carpet Tiles Machine (Pahl and Beitz 1996: 33)*




representation of an intended behavior that is to become an artifact's function and the sub-behaviors that compose it (and that are to become sub-functions of the artifact's sub-systems and components). Context will make clear which of the two is meant, but when the difference is important I will be more explicit.

## 4.2 Examples of Functional Decompositions

To get a handle on our topic we start with the empirical facts. In our case, they are what engineers and design methodologists say about functional decomposition and its uses. We will look at a few examples from design methodology textbooks in this section and focus on some more advanced treatments in journal articles in the next section. After every example, I will point out salient features that warrant further attention.

### *4.2.1 Stamping Out and Packing Carpet Tiles*

Many engineering textbooks stress the importance of functional decomposition. The description they give of the process is typically pretty straightforward: 'A complex *overall function* can be broken down into *sub-functions* of lower complexity' (Pahl and Beitz 1996: 150). Pahl and Beitz take a function to be an input-output relation of a system that is supposed to perform a task. (Note that, in my terminology, this is exactly a behavior.) The point of functional decomposition is to produce what they dub a *function structure*: a breakdown of an overall function into main sub-functions and auxiliary sub-functions. The former contribute directly to the overall function, the latter indirectly. Once this has been done, the various sub-functions must

be translated into physical processes, characterized by their physical effects. By way of example, they suggest that a function like amplification could be translated into different physical effects, such as the lever effect, wedge effect, hydraulic effect, or electro-magnetic effect (ibid.: 38–40). These effects can then easily be translated into specific physical processes. They also provide various illustrations of function structures. Figure 1 shows the function structure of a machine that stamps carpet tiles out of a length of carpet and packs them. The main sub-functions are depicted in a row and the auxiliary sub-functions above and under this row.

The example may seem rather straightforward, but I think there is more to it than meets the eye. Let us look at some of its features. For the sake of clarity I will label the features with capitalized letters so we can easily refer back and expand on them in later examples.

(A) Note how the decomposition not only carves up the overall behavior, but also interprets and specifies that behavior further. Or perhaps I should say that it makes implicit assumptions about how the overall behavior is to be carried out explicit. The overall behavior of 'stamping out and packing carpet tiles' does not imply sub-behaviors like 'check quality' or 'remove off-cuts.' These sub-behaviors derive from quality standards that are presumably taken for granted and implicitly understood common practices. What this means is that, even though overall device functions are often described very briefly and simplistically, they often presuppose unstated specifications that are made explicit in a functional decomposition. Spelling those out may be a lengthy process that requires a lot of expertise about the relevant product, design practice, and legal or other norms.

(B) The identified sub-behaviors are not ordered arbitrarily. They form a fairly rigid chronological order; most changes would lead to serious problems or impossibilities. In fact, the identified main sub-behaviors are sequential steps to accomplish the desired result; the first has to be finished before the second can start, etc. The auxiliary sub-behaviors must also conform to this temporal sequence; off-cuts can only be removed after they have been cut off and separated, the packing material has to be supplied after the combining of the individual tiles into lots so that the packing can start. The moral is that, at least in some cases, functional decomposition orders the sub-behaviors in time.

(C) The sub-behaviors are also ordered in space. It is silently assumed that the events that occur when these sub-behaviors are carried out take place at the same location or adjacent locations. If the cutting were to happen at another place than where the carpet is provided, there would never be any tiles. This is so obvious that it easy to see why it should not be made explicit. A functional decomposition, then, also includes or else implies information about the spatial

ordering of the behaviors taking place. Perhaps it is more accurate to say that it implies information about the locations of the objects participating in the performance of the sub-behaviors. This information need not be complete or maximally specific, but it does contain rough indications of the relative spatial positions of the participating objects and, thereby, the events taking place when the sub-functions are performed.

(D) It is often urged that functional decomposition ought to be done before any structural characteristics of the device being designed are introduced and that it must be done independently of structural characteristics (see also section 4.2.3 below for this). Since ought implies can, those giving this advice must be assuming that functional decomposition *can* be done independently of any ideas about the artifact's structure. Let us give this assumption a name, *Independence of Structure* or (IS) for short, because we will need to talk more about it.

> (IS) Functional decomposition can be done independently of any ideas about the structure of the artifact that is being designed.

The example under discussion does not clearly live up to this standard. The given decomposition has built into it a number of presuppositions about the device performing the overall behavior. The most general one is that all the sub-behaviors identified are to be performed by one machine; Pahl and Beitz say it is a functional decomposition for a (singular) machine. This need not be so. One can easily imagine set-ups in which one machine does the cutting, another the grouping of tiles, and yet another the packing. Another presupposition is that the tiles are *stamped out*, as opposed to, say, cut out or torn. Examples could be multiplied, but the point is that this functional decomposition presupposes a number of things about the machine that is supposed to be performing the overall behavior. This seems to be in tension with (IS). This is not supposed to be a criticism, but it is a significant point that we should come back to later.

### *4.2.2 Coffee Grinder*

Roozenburg and Eekels, who call the technique function analysis, illustrate a variant of Pahl and Beitz's approach by means of the function of a coffee grinder (1995: 194–199). Figure 2 shows their example. Some of the features I discussed above are also present here.

(A) It is even more obvious that the initial description of the overall behavior is fairly unspecific. The device's behavior is not merely to grind the coffee beans (note how grinding beans is in fact listed as one of the sub-behaviors, while it was also identified as the overall behavior), but it should allow for different degrees of coarseness and be able to store some amount of beans and ground coffee.

***Figure 2*** *Function Analysis of a Coffee Grinder (Roozenburg and Eekels 1995: 199)*

| | |
|---|---|
| Grind coffee beans | (Transforming matter) |
| Adjust grinding coarseness | (Connecting matter and information) |
| Supply grinding energy | (Connecting matter and energy) |
| Switch on/off | (Connecting energy and information) |
| Turn electric energy into mechanical energy | (Transforming energy) |
| Store fresh and ground beans | (Storing matter) |

(B) A remarkable feature of this example is that the sub-behaviors for grinding coffee are not ordered chronologically. Sub-behaviors seem to be simply listed in no particular order.

(C) This decomposition also implies information about the spatial locations of the various behaviors. For example, the conversion of electric energy into mechanical energy, the supplying of grinding energy, and the actual grinding will have to occur at roughly the same spot.

(D) Here, too, the (IS) advice is not strictly obeyed. The decomposition seems to presuppose at least a partial idea of what the prospective coffee grinder should look like. For example, it will be powered by electricity, not by human kinetic energy and it has an on/off switch instead of merely a power cord and plug.

(E) This example also shows something that was not so clearly present in the previous example. Functional decomposition can involve describing behaviors at higher or lower levels of abstractness. A description like 'turn electric energy into mechanical energy' is so abstract that it could count as a sub-behavior of nearly all modern electronic equipment. But the details of this sub-behavior for the coffee grinder could also be described much more concretely, by specifying the exact voltages and currents and the specific mechanical movements and forces. Something like this happens rather implicitly in the third and fifth sub-behaviors, where the very abstract 'mechanical energy' is specified by calling it 'grinding energy'.

### *4.2.3 Splashgard*

David Ullman, in another design methodology classic, defines a function as the behavior of a machine or human that is necessary to fulfill the design requirements. To introduce functional decomposition, he simply states that the

overall function has to be decomposed into sub-functions (1992: 142). He does provide some additional guidelines for doing that properly. Not all of them are relevant for the present purposes, but those that are, read as follows (ibid.: 144):

(a) Document what the artifact needs to do, not how it needs to do it;
(b) Match inputs and outputs in the functional decomposition;
(c) Break the function down as fine as possible.

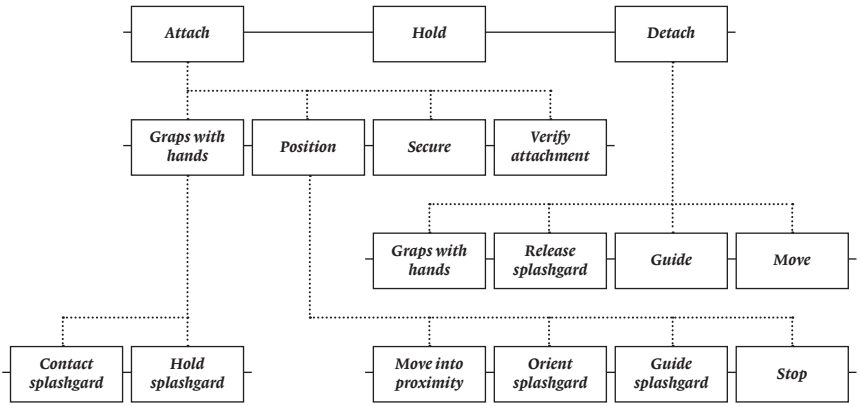
(D) We can note in passing that the first guideline confirms that design methodologists assume (IS). 'What the artifact needs to do' presumably refers to sub-behaviors, while 'how it needs to do it' must stand for structural characteristics.

(F) The third guideline is somewhat vague as long as the relevant sense of possible is not described in more detail. I am sure that an experienced car mechanic will be able to break down the function of my car much more finely than I will ever be able to, but presumably a trained mechanical engineer could go even farther. A materials engineer may then be able to pick up on some of the sub-behaviors that the mechanical engineer has identified and decompose these further in terms of the behaviors of certain types of conglomerates of molecules. After that, chemists and physicists could perhaps decompose the behaviors of these molecules into the behaviors of atoms and even more basic particles. It would be absurd if the third guideline were interpreted as requiring that every functional decomposition end in fundamental physics. The relevant sense of 'as fine as possible' must be more pragmatic. It seems plausible that certain sub-behaviors will count as basic or atomic for particular engineering disciplines. Which will vary across different fields; where mechanical engineering will accept, say, the strength of a particular alloy as given, materials engineering will be interested in how this strength is realized by the behaviors of elements in the alloy.

Ullman goes on to give an example of a *Splashgard*, the function of which is to deflect water. Figure 3 shows how he proposes to break down its overall behavior into several layers of sub-behaviors.

(G) At first sight, the example may seem similar to the other ones, but when we look closely this decomposition is rather different from those we looked at before. The sub-behaviors it identifies are not so much behaviors of the Splashgard as they are actions of someone using the device: grasping, securing, positioning, holding, guiding, etc. are all things you are supposed to do *with* it, as opposed to things the Splashgard itself does. In fact, this decomposition looks more like a use plan for the Splashgard than a functional decomposition of its behavior. A better name for it would perhaps be 'action decomposition.' To the extent that this example is representative for a larger class of

***Figure 3*** *Functional Decomposition of the Splashgard (Ullman 1992: 145–147)*




functional decompositions, then, functional decomposition is apparently also about identification of actions that involve manipulation of an artifact. That process, however, is the province of use plan design and as such it falls outside the limitations of my project in this thesis as I said in chapters 1 and 2. I acknowledge that this type of 'functional' decomposition is important in engineering design, but I will not analyze it any further.

### *4.2.4 Disposable Razor Blades*

Finally, in a classic in the design methodology field, J. Christopher Jones (1970: 331–340) describes a method he calls *Functional Innovation*, which also involves the gist of functional decomposition, but now of the function(s) of an already existing product. The method has five steps:

(1) Identify the functions of each physical component of the existing artifact.
(2) Identify essential functions of which these are sub-functions.
(3) Identify any changes to the essential function that are likely to be beneficial.
(4) Combine (2) and (3) to give a revised essential function.
(5) Seek alternative ways of subdividing the revised essential function and of allocating each sub-function to a new but feasible physical component.

Steps (1) and (2) identify the functional decomposition of the function(s) of an existing artifact. It seems to me that steps (4) and (5) are very similar to the other methods we have looked at, except that step (5) explicitly includes an

***Figure 4*** *Functional Innovation of the Function of Razors (Jones 1970: 332)*

| · **Existing physical components** | · **Existing functions** |
|---|---|
| Blade | Control |
| Handle | Cutting |
| Strop | Sharpening |
| · **New physical components** | · **New sub-functions** |
| Blade edge | Cutting |
| Blade clamp | Stabilizing cutting element |
| Blade clamp | Transmitting of force to cutting element |
| Blade clamp | Control of cutting angle |
| Handle | Control of strokes |
| Handle | Transmitting force from hand |
| Replacement of blade | Maintaining sharpness |

allocation of the sub-functions to physical components, whereas the previous methods claimed to be independent of physical implementation. This method overtly depends on pre-existing artifacts and the way in which they and their components fulfill particular functions and sub-functions. Its point of departure is an existing device and it seeks to improve what that device does by doing more things, other things, or the same things better.

(D) This version of the method immediately and unequivocally explicitly takes into account what is physically feasible. Although Jones says nothing about how the decomposition of the new function and the allocation to physical components are to supposed interact or constrain one another, he does at least show some sensitivity to the fact that functional decomposition may be more closely related to structural implementation than those who assume (IS) will admit. As an example of this method Jones discusses disposable razor blades. Figure 4 depicts his idea by enumerating the old and new components with their various sub-functions.

A bit more about feature (D), the alleged independence of functional decomposition from structural considerations: For each example I have pointed out dependencies between the functional decomposition and structural details of the artifact implementing the function. Because Jones's method is so clear about its dependence on pre-existing artifacts, his example brings out nicely how functional decomposition may be driven by considerations about the structure that is going to implement the overall behavior. At the same time, however, it shows that some processes of functional decomposition may indeed be relatively independent of structural considerations. To see this, note how the example can be interpreted in two ways: structure-driven and behavior-driven.

On the one hand, the invention may have developed out of an idea about a new physical structure in which the blade is no longer of one part with the handle, but instead easily replaceable. I am not saying that this is in fact how the idea arose but it does seem to be a live possibility. On the other hand, someone may have been thinking about the behavior of maintaining sharpness and come up with the invention of replacing blunting parts, regardless of any particular structure. This new behavior would then have necessitated a different structure than before. Which of these two scenarios really happened is unimportant. The point that deserves emphasis is that, in contrast to (IS), functional decomposition is not always wholly independent of structural considerations. The authors we have looked at do not seem to be sensitive to this point and as a result they have little to say about the exact type of dependency at issue. This is something we will need to look into more deeply.

### *4.2.5 Lessons About Functional Decomposition*

I think it will be helpful to summarize the central features of functional decompositions that we have discovered in our exploration of design methodology textbooks. These features set the agenda for the rest of the chapter.

(A) Functional decomposition specifies the overall desired behavior and makes suppressed assumptions about it explicit. Since overall behaviors are often characterized incompletely by very succinct descriptions, it is only through a complete functional decomposition that one comes to learn the precise desired behavior of artifacts. It is evident that this process of specification is a *conditio sine qua non* for a successful design process. The artifact's intended behavior must be specified in every relevant detail in order to find an appropriate structure for the artifact. It seems to me that this is pretty much self-evident, so I will not have much more to say about it.

(B) & (C) Functional decompositions often include implicit or explicit information about the temporal and spatial order of the sub-behaviors. Some functional decompositions can be read as timelines charting sequences of required events. The locations where the sub-behaviors are to be performed form an implicit backdrop to functional decompositions, as it is assumed that the sub-behaviors are carried out at identical, adjacent, or otherwise relevantly related locations. As spatial information is structural information, a process of functional decomposition already constrains part of the artifact's structure. As long as the components that will carry out the sub-behaviors are not specified in further detail, however, these constraints are very minimal; only relative locations of the components, whatever their structure will be, within the artifact are fixed.

(D) The previous point established that functional decomposition constrains

the structure of the artifact in a minimal sense, but we have also seen that many functional decompositions have stronger commitments because of the specific sub-behaviors they identify and the particular ways in which they carve up an overall behavior in sub-behaviors. Some functional decompositions have their prospective structures written all over them, so to speak. They do not entail one unique structural description of an artifact, but anyone with some familiarity with the sort of artifact being designed will easily see the basic elements of its prospective structure. All of this has to do with assumption (IS). It is too soon to draw definite conclusions here, but it is fairly clear that there are reasons to doubt (IS). We should definitely delve deeper into the relation between functional decompositions and structural descriptions of artifacts.

(E) Some functional decompositions move from more abstract and general descriptions of behaviors to more concrete and specific descriptions of sub-behaviors. As we saw in the first chapter, descriptions of desired behaviors that are the starting point of an artifact design process can be highly abstract and part of the challenge of design is to translate this abstract description into a more concrete one. This feature of functional decomposition thus gives us reason to believe it contributes to that task. Again something we will need to get back to.

(F) A process of functional decomposition stops when the identified sub-behaviors are considered to be sufficiently basic according to relevant standards. These standards are relative to engineering disciplines and perhaps even to design situations. What counts as basic for a mechanical engineer need not be basic for a materials engineer, and where most normal design situations will take the behaviors of standardized components as sufficiently basic, other innovative design situations may require that even these sub-behaviors are decomposed further. The important point is that the transition from a dispositional description to a categorical description of something realizing it will be a lot easier for basic behaviors than for non-basic behaviors. That is because it is part of the notion of a basic behavior that designers are familiar with it and therefore know what sort of things have the right dispositions. Yet another topic that deserves further attention.

(G) Finally, some authors include not only behaviors of the artifact itself in their functional decompositions, but also actions of people using the artifact. To the extent that this is common practice, functional decomposition is not only about what the artifact itself does, but also about what its users do with it and thereby about use plan design. This falls outside the scope of my project, so I will say nothing more about it here.

Before we start to analyze these features in more detail, I will use the next section to look at some work that builds on the sketchy accounts of functional decomposition from this section.

## 4.3 Further Developments in Functional Decomposition

The fairly unsophisticated textbook accounts from the previous section have spawned a voluminous body of research papers on themes related to functional decomposition. A lot of work in this area is driven by practical interests (cf. Stone and Wood 2000; Hirtz *et al.* 2002; Chandrasekaran 2005). First, many designing engineers felt the need for a clear stopping point in the process of functional decomposition; 'atomic' sub-behaviors that need not or cannot be decomposed further. These would have to be described at a consistent level of detail, so they thought. Second, for the purpose of more effective and efficient communication in design it also seemed important to have a common vocabulary to describe artifacts and their functions and sub-functions. A related benefit of this would be the possibility of creating some sort of measure to compare different products with regard to their function. A third motivation came from an interest in representing (design) knowledge about artifacts and their functions in database computer systems. Such an effort, too, requires a common vocabulary and in addition an unambiguous hierarchical ordering of behaviors and sub-behaviors. Fourth, as we saw in the previous chapter, design methodologists with an interest in artificial intelligence (and artificial intelligence people with an interest in design) have been trying to develop intelligent agents that can take care of specific tasks in engineering design processes or, more ambitiously, carry out complete design processes all by themselves. In this context, it also seemed crucial to have a univocal vocabulary for describing the functions and behaviors of artifacts.

The bulk of this work does not aim directly at improving or refining the method of functional decomposition. It focuses on related issues: how to understand behaviors and functions, sub-behaviors and sub-functions, and the relations between them. Broadly, there are two main streams in the field. The first concentrates on the representation of functions and typically goes under the heading of *functional representation* research.[33] The second concentrates on adequate modeling of functions and one of its most important goals is to find a *functional basis*, i.e., a complete set of atomic behaviors that would suffice to describe any given device in terms of its behaviors. This research is often referred to as *functional modeling* research.

### *4.3.1 Functional Representation*

In functional representation research a lot of effort has gone into the development of precise notions of function, behavior, object, and structure.

[33] This research is thus more related to the development of an adequate function theory than to understanding what functional decomposition is. The reason why I do include it here is that it touches on issues that are relevant for functional decomposition.

One influential paper in this area (Chandrasekaran and Josephson 2000) distinguishes between two different meanings of function: (1) function as effect on the environment and (2) function as what a device does. Common to both meanings is that an artifact's function is something it does or, more precisely, a desired or intended behavior of the artifact. This behavior can be characterized differently, giving rise to the two different meanings: first, in an environment-centric way that focuses on what an artifact accomplishes in its environment, and, second, in a device-centric way that focuses on the artifact's properties or behavior without explicit mention of how they affect the artifact's environment. The first description includes a mode of deployment, i.e., a way in which an artifact is to be employed in a relevant environment. It specifies the relevant causal interactions between the artifact and its environment. What an artifact accomplishes when used as stipulated by a mode of deployment, i.e., a set of behaviors, constitutes what Chandrasekaran and Josephson call a role. A function, then, is an intended role. The second notion of function suppresses explicit references to an environment or interactions with it; they are implicitly assumed. Device-centered functional descriptions just consist of a set of desired or intended behaviors that are specified purely in terms of what the artifact itself does. Although this may require inputs from the environment, these inputs appear in the description only as conditions on behaviors of the artifact itself.

Here is an example of what Chandrasekaran and Josephson have in mind. They talk about an electrical buzzer composed of two boxes, *box 1* and *box 2*, connected by a wire. *Box 1* has a switch. The function-as-effect of this buzzer is 'to provide a means by which a person at *location 1* may cause sound to be produced at another location.' Its function-as-what-a-device-does is 'to make a sound come from *box 2*, when the switch in *box 1* is closed.' The difference between the two descriptions, Chandrasekaran and Josephson contend, is that the former describes the desired behavior 'entirely in terms of elements external to the device,' while the latter does so 'entirely in terms of the device.' (ibid.: 170)

A few brief remarks about this. First, it is not quite clear to me that these two construals of function in fact constitute two significantly different meanings of 'function.' It rather seems that function-as-effect is a more general way of describing what an artifact does, which abstracts away more from how it does what it does. Function-as-what-a-device-does is more concrete and gives more how-information by describing the artifact in terms of behaviors of its actual components, which inevitably involves some suggestions – not necessarily complete in any sense – as to how it does what it does. If Chandrasekaran and Josephson want to insist on calling these two meanings fundamentally different,[34] they should concede many more meanings; one for each additional description of the artifact's behaviors at another level of abstraction. We could

easily add to the buzzer example another more abstract description of the buzzer's function that says even less about how it does what it does by omitting the fact that the buzzer makes a sound to perform its function. The function could be described as 'to enable a person in one place to alert another person somewhere else.' The same goes for the more concrete end of the spectrum. The buzzer's function could be described as 'to close an electrical circuit so that a current starts running, which activates a little sound system and loudspeaker that produces a high-pitched tone.' I think the bifurcated picture should be replaced by one in which descriptions of functions can range from very concrete to highly abstract; the more abstract the description, the less specific details about the artifact itself are given. This harks back to the point about dispositional and categorical descriptions and different levels of abstraction from the previous section, feature (E). I will come back to this in due course.

Second, as a theory of technical functions this account is too liberal in that it suggests that an intended role suffices for a function: 'Functions are Roles + Intentions' (ibid.: 172). Mere intentions do not suffice to bestow artifacts with functions as we saw in chapter 2. An intention to ascribe a technical function to an artifact must be based on beliefs with positive epistemic status about the artifact's dispositions. This remark may be beside the point, however, since I am not sure whether the authors really want their account to capture the everyday or engineering notion of technical function; they may have some other more modest purpose in mind.

### *4.3.2 Functional Modeling*

Now for the functional modeling research. One of the main thrusts there is to find a *functional basis*: a set of functional primitives or atomic behaviors that would suffice to describe the behavior of any given device in terms of, ultimately, atomic behaviors.[35] As can be expected for such an ambitious goal, people have come up with different proposals that are not always easy to reconcile. For example, Pahl and Beitz (1996: 36) list five generally valid behaviors: changing of input/output type, varying of input/output magnitude, connecting of input/output, channeling (i.e., change of input/output place), and storing (time of input differs from time of output). The input and output for these behaviors consist of flows of matter, energy or information. Many

[34] I am not sure how strong the distinction between the two meanings is supposed to be. Although the authors admit the possibility of mixed forms, they also talk about these two meanings as being representative of two basic classes of descriptions of functions.

[35] The authors talk about functions, but since functions are intended behaviors and since we are investigating functional decomposition as it is used in *design* contexts here, I interpret their use of 'functions' as being about *behaviors* that are to become functions.

subsequent proposals refine and expand on this abstract list. A complete overview of these proposals would lead us too far astray. I will summarize the main ideas of one of the most developed attempts in this field, expounded in Stone and Wood (2000) and further refined in Hirtz *et al*. (2002). The three basic flows identified by Pahl and Beitz are re-identified as three basic *classes* of flows: material, energy, and signal. In each class, Stone and Wood distinguish several basic flows and some of these basic flows are further subdivided into sub-basic flows. For example, the class of energy contains the basic flows human energy, acoustic energy, biological energy, electrical energy, electromagnetic energy, etc. The basic flow electromagnetic energy is then further subdivided into the sub-basic flows of optical and solar electromagnetic energy. All of these flows can figure as inputs and outputs for behaviors of artifacts and their components. A similar approach is taken with behaviors themselves. The authors distinguish eight very general classes of behaviors: branch, channel, connect, control magnitude, convert, provision, signal, and support. Each class has at least two basic behaviors in it. For example, the class branch contains separate and distribute as its basic behaviors. The basic behavior separate is further subdivided into divide, extract, and remove. The authors even provide lists of synonyms for all of their behaviors, to make sure that people do not confuse synonyms with additional levels of specificity. The total list of basic behaviors contains over 20 items, not counting further subdivisions. The authors are confident that they are on their way to a complete set of truly atomic behaviors: 'these sets of terms could be distilled to a more fundamental set that would ideally … lead to a minimal set of terms that did not overlap, and yet provided complete coverage of the space of designed products' (Hirtz *et al*. 2002: 75).

Again a few remarks. First, a brief point of clarification: In contrast with the functional representation research, the notion of function used here should not immediately be understood in the stronger intentional sense of a technical function. I took this into account in my summary by consistently talking about behavior and not about function.[36] The functional modelers' sense of 'function' is more akin to that of 'role' in Chandrasekaran and Josephson (2000); it can be used for anything artifacts or their components do, and not just for the intended behavior that is its technical function. I should add immediately, however, that since technical functions are intended behaviors or dispositions (see chapter 2 for details on this), the vocabulary is perfectly suitable to describe technical functions as well. In either case it describes them at a low level of abstraction; it gives concrete details about the behaviors and their inputs and inputs.

Second, many a philosopher may be tempted to start quibbling about this list. Do the items on it really exhaust the space of artifact behaviors and are

[36] Cf. also note 35 above.

they mutually exclusive? Are some reducible to others? Can they be rearranged according to higher-order organization principles? Such an undertaking may have some merit but I will not venture into it here. My aim is not to improve on anyone's list of atomic behaviors, but to understand functional decomposition and the possible relations between behaviors and sub-behaviors better.

Third, another observation about the relation between dispositional and categorical descriptions and features (B) through (D) from the previous section. The atomic behaviors identified in this research do not seem to have any categorical implications at first sight. It looks as if they form a purely dispositional vocabulary, fully independent of categorical information. I think this impression is right in one sense, but wrong in another. The atomic behaviors identified are indeed fairly abstract and, considered in isolation, imply little to nothing about the structures implementing them. Nonetheless, once they are used in functional decompositions they do confer categorical information. First, the descriptions of particular sub-behaviors with their inputs and outputs in the vocabulary provided by a functional basis are fairly detailed. Because of this specificity, they imply categorical information about how the artifact will realize that function. For example, if you see the sub-function 'convert solar energy to electric energy,' that tells you something about the sort of power supply that the device at hand uses. Second, the choice to break up a given behavior in particular sub-behaviors rather than others also implies a choice for one particular kind of structure rather than another. One of the sub-behaviors of a computer mouse is to convert kinetic energy into a digital signal that controls the movements of a mouse pointer on a computer screen. Wheel mice do so by converting the movements of a little rubber ball into an electric signal but optical mice translate changes in a light signal into an electric signal. The different atomic sub-behaviors required for this tell you something about the different categorical details of computer mice. So again we see that descriptions of behaviors and functional decompositions tend to imply structural information. The point is still somewhat imprecise; we shall need to look closer at how this happens exactly.

Fourth, suppose there truly is a functional basis in the sense intended by the functional modeling people and also suppose that Hirtz *et al.*'s list is a pretty good approximation of it. In what sense are the behaviors in this set atomic? Not in the sense of being the stopping point of functional decompositions. The goal of functional modeling is not to identify such a stopping point, but to form a common vocabulary in which all artifacts can be described in terms of their behaviors. No one says functional decomposition must always continue until it hits the rock bottom of atomic behaviors. So 'atomic' does not have the same sort of relativity to engineering fields and design tasks as the stopping point of functional decomposition has (see feature (F) in the previous section).

Nonetheless, it still seems to me that what counts as atomic is relative to engineering fields in a broader sense. Most functional modeling people come from mechanical engineering or similar backgrounds and the atomic behaviors they identify are typical for mechanical engineering and fields that work with artifacts in the same order of size and complexity, such as civil engineering, aerospace engineering, and parts of electrical engineering. There is nothing wrong with that, but engineers in other fields, e.g., materials engineers, some electrical engineers, and nano-engineers, will probably need a vocabulary with behaviors that are even more fine-grained. Their atomic behaviors must include the typical behaviors of (compounds of) molecules, atoms, or even smaller fundamental particles. But since I am not primarily interested in the merits of functional modeling research, we can put this point aside. What should be kept in mind for the present purposes is that claims about basic or atomic behaviors tend to be relative to unstated assumptions about engineering fields, the sort of design task, and other epistemic or pragmatic factors. Making such assumptions explicit is important in order to understand functional decomposition better.

### *4.3.3 Concluding Remarks*

This concludes my review of the literature related to functional decomposition. The rest of the chapter is devoted to analysis of this literature and the points it brought up. We have a number of issues on the table. The first has to do with the relation between dispositional and categorical descriptions of artifacts and their components. In the examples we looked at in section 4.2, we saw how descriptions or behavior, especially when they are ordered in a functional decomposition, often imply bits of categorical information about the artifact and its components (features (B) through (D)). The same thing happens in the work of the functional modeling people as we just noticed. In order to understand functional decomposition better, we need to investigate how descriptions of behavior and functional decompositions imply categorical information. The second issue that surfaced both in the previous section (feature (E)) and in the discussion of functional representation research above is that behaviors and sub-behaviors can be described at various levels of abstractness. Some sub-behaviors are more concrete descriptions of the behavior(s) they compose. The third issue that needs to be sorted out is the stopping point of a process of functional decomposition. This was feature (F) in the previous section and the issue cropped up again in relation to the atomic behaviors in the functional basis that the functional modeling people are looking for.

In the rest of the chapter, I will develop an analysis of functional decomposition that reproduces and explains these features. The first step,

to be taken in the next section, is to show how the dispositional-categorical distinction as introduced in chapter 2 already accounts for a number of these features. After that, we have to face a more fundamental question that we have not touched upon yet but that looms large: What is (are) the exact relation(s) between a behavior and its sub-behaviors? An answer to that question will help us truly see what happens in a process of functional decomposition.

### 4.4 Dispositional and Categorical Descriptions of Artifacts

To a greater or lesser extent, the authors we have been looking at seem to be committed to assumption (IS), which says that functional decomposition can be done independently of any ideas about the prospective artifact's structure. Presumably, this is one reason why Pahl and Beitz set the transition from sub-behaviors to physical processes apart as an additional step, after the basic functional decomposition has been laid out. Most people do not state the assumption explicitly; it is lurking in the background only to be read between the lines. An exception to this rule, business information systems engineer Jan Dietz explicitly assents to (IS). '[A functional decomposition] is a purely conceptual division of the function or functionality of a system, independent of its construction and operation' (2006: 69). David Ullman's first guideline in section 4.2.3 ('document *what* the artifact needs to do, not *how* it needs to do it') is also pretty explicit.

Assumption (IS) indeed is initially plausible. For one thing, a functional decomposition of a given artifact does not entail a unique categorical description of that artifact and that suggests that functional decompositions indeed are independent of categorical descriptions in some sense. Another source of seeming support for (IS) comes from my own endorsement of a sharp distinction between dispositional and categorical descriptions of artifacts. On Mumford's view of that distinction there are no entailments from dispositional descriptions of an object to categorical descriptions of it. If it is added to this that functional decompositions are indeed (purely) dispositional descriptions of artifacts, (IS) seems to follow directly.

Nonetheless, the main point of this section will be to argue against (IS) or, more precisely, to argue that (IS) gets at an important point but is too crude in its current formulation. The previous sections already prepared the way for this by hinting at problems with (IS), but here I will pinpoint those problems, explain them, and propose a better view of the relation between a functional decomposition of an artifact's intended behavior and a categorical description of an artifact. A careful application of the distinction between dispositional and categorical descriptions of artifacts will be the key to my analysis.

One caveat before we continue: Perhaps someone comes off with the impression that I am knocking design methodologists for endorsing (IS), while

I was the one who suggested they do so in the first place. In other words, I could be accused of committing a straw man fallacy. My reply is that I would be happy to learn that design methodologists in fact deny (IS). I have come to believe they endorse something like it and I have presented my evidence for this belief, but of course I might be mistaken. If so, all the better. In the end, my goal is not to criticize design methodology but to assist it. If design methodologists embrace my corrections to (IS) as being what they really had in mind all along, I would consider my goal realized. My only remaining quibble would then be that I would have wished they had been a little more explicit about their views.

### *4.4.1 The Dispositional-Categorical Distinction Revisited*

Design methodologists often use the locutions 'what an artifact does' and 'how it does what it does' to distinguish – I take it – dispositional descriptions of an artifact's behavior from categorical descriptions of its structure. So far in this chapter I have followed suit but this is unacceptably imprecise from a philosophical point of view. I propose we adopt Mumford's account of the distinction between dispositional and categorical descriptions. (We should not proceed to assume that functional decompositions as a whole are good examples of purely dispositional descriptions, but we will get to that later.) If the legitimacy of interpreting what vs. how as dispositional vs. categorical is not immediately obvious, recall that design methodologists talk about what an artifact does and what its components do in terms of verbs or input-output relations. In saying what an object does they describe it in terms of its response to being used or stimulated in some way. In other words, they ascribe behaviors to objects. In doing so they ascribe properties to objects, namely such properties that make it the case that the object can show the behaviors ascribed to it. These properties are responsible for the object's behaving in the way it does when it is stimulated in the way it is. Obviously, that is not saying much about these properties themselves except that the object must actually have them and that they are such that they lead to this particular behavior manifestation when the object is stimulated in such-and-so ways. All of this is perfectly in line with what I said in chapter 2 about dispositional descriptions. Dispositional descriptions are connected with subjunctive conditionals by conceptual necessity. To describe an object dispositionally is to say that, under context-determined ideal conditions, were it to be $\chi$-ed, it would $\phi$.

For the sake of completeness, let me also briefly repeat what I said about categorical talk. Categorical descriptions have no conceptual connection with behavior manifestations. They only specify shapes, structures, and states: things like triangularity, molecular structure, atomic number, length, temperature, etc. In doing so, they also ascribes properties to objects, but these properties are not characterized in terms of the behaviors they make

possible. Categorical descriptions entail the existence of properties, but, unlike dispositional ones, they do not analytically entail anything about what sort of behaviors the object having these properties can fulfill.

On Mumford's view, then, the distinction between dispositional and categorical is a distinction at the epistemic level of our concepts and not an ontological reality. At the ontological level, there is only one sort of properties. It is okay to talk about dispositional and categorical properties but what that means, strictly speaking, is properties-dispositionally-described and properties-categorically-described. For example, the properties underlying the stiffness of a piece of metal, which consists in its showing characteristic behaviors when certain forces are applied to it can also be described categorically as shapes and molecular structures. At the same time, the molecular structure of the metal will also underlie the (lack of) transparency of the metal, which is another behavioral role of that same property. Another example: the stickiness of Velcro can be accounted for in terms of geometrical shapes of the loops and hooks on pieces of Velcro. These same geometrical shapes are also behind the prickly sensations you get from pressing your finger on a piece of Velcro.

### *4.4.2 Applying the Dispositional-Categorical Distinction*

This was familiar terrain for anyone who read through chapter 2. I now want to use the conceptual apparatus of the dispositional-categorical distinction to reconsider (IS). First a small terminological point. With the present distinction in hand we can see an ambiguity in the original design methodologists' distinction between 'what an artifact does' and 'how it does what it does.' Describing how an artifact does what it does can be done by means of both dispositional and categorical descriptions. For instance, to explain *how* a light bulb lights a room you could describe the behaviors of the light switch, the glass bulb, the gaseous mix inside, and the filament. Surely, that would not be a categorical description. In fact, functional decompositions themselves can be interpreted as answers to how-questions in this sense because they show how artifacts realize their overall behavior by manifesting various sub-behaviors. (This fact renders Ullman's guideline to 'describe what an artifact needs to do, not how it needs to do it' rather problematic.) At the same time, describing how an artifact does what it does can also be done by means of categorical descriptions. The filament in a light bulb glows, ultimately, because of its shape and molecular structure. A description of those might also figure as a description of how the filament does what it does.

The second point is that the present distinction implies a claim that is subtly weaker than (IS), but not (IS) itself. When dispositional and categorical descriptions are distinguished as I have proposed, singular dispositional

descriptions do not entail specific categorical descriptions. Knowing that an artifact would $\phi$ were it to be $\chi$-ed does not tell you anything specific about, say, its molecular structure or some other categorical fact.[37] Not, at least, without further background knowledge as we will see shortly.

The point works analogously the other way around: Singular categorical descriptions do not entail specific dispositional descriptions. This has been a bone of contention in discussions about dispositions. In particular, D.H. Mellor (1974; 1982) has urged that categorical property ascriptions entail conditionals about an object's behavior in the exact same way as do disposition ascriptions. His example is that the description of an object as triangular analytically entails that, if someone were to count its corners correctly, the result would be three. One may question whether this is a dispositional description *of the object* at all, but let us grant Mellor that point. I think Mumford's reply (1998, §4.6) is convincing and I will summarize it here. Mellor is right to point out that it is analytical that correctly counting the corners of a triangular object gives the result three, says Mumford. That is simply what it means to count correctly. What is not analytical, however, is that we are in fact able to count corners correctly. One can easily imagine possible worlds in which we are systematically deceived about the number or corners objects have. For sure, the actual world is not such a world, but that is beside the point since analyticity requires truth in all possible worlds. Most of us are pretty reliable corner-counters, to say the least, but that ability does implicitly depend on the contingencies of how our world works. And so it is in general; categorical descriptions of an object do not analytically entail anything specific about (possible) dispositional descriptions of that object. The fact that it is very easy for us to deduce one or more possible behaviors once we know some of an object's categorical descriptions does not affect this point, because such derivations can only be made *given* the vast body of everyday experiential knowledge we all have or perhaps the specialized knowledge that some of us come equipped with after years of formal education.

Combining the points from the last two paragraphs yields the following claim: Describing an object dispositionally can be done independently of beliefs about its categorical structure. This claim sounds like (IS) but its scope is less impressive. It is about a dispositional description of an object as opposed to a functional decomposition of it. This matters because a functional

[37] Nothing *specific* because you could argue that it does tell you that the thing has some molecular structure or other. Or that it must have a molecular structure that is such that it enables the described behavior. But these pieces of information are gratuitous because it is part of the notion of a material object to have some molecular structure or other and the fact that the molecular structure is such that it enables the described behavior follows trivially from the dispositional description and does not contain specific new information about this molecular structure.

decomposition consists of many dispositional descriptions that are ordered in specific ways. What holds for isolated dispositional descriptions does not automatically carry over to functional decompositions.

Third, the move from singular dispositional descriptions to functional decompositions changes things. A functional decomposition, as we saw in the examples above, is a set of dispositional descriptions that typically has a spatiotemporal order imposed on it. I now want to show that categorical information is implicated in functional decompositions in at least three ways.

(a) The most obvious way is through the (sometimes implicit) spatiotemporal ordering of the described behaviors. This ordering constrains where and when the behaviors take place and that says something about where the components that are to take care of them must be located relative to one another. Clearly this is not saying a whole lot as long as the behaviors have not been allocated to (combinations of) components and nothing further is known about the shapes and material constitution of the components, but it is categorical information nonetheless. In Pahl and Beitz' carpet tiles machine, for instance, the component(s) responsible for the stamping of tiles must be adjacent to that (those) taking care of the quality checking. If the relative times of behavior manifestations also count as a sort of categorical information, functional decompositions that are chronologically ordered provide this information. Bottom line: Functional decompositions imply categorical information about relative locations of the artifact's physical components. Nothing further is implied about the number of components, and their shapes, physicochemical makeup, or other categorical characteristics; all of that has to be sorted out later.

(b) Functional decompositions may also imply categorical information through the specifications of the sub-behaviors and their inputs and outputs. This is, of course, conditional on whether or not and how detailed inputs and outputs are specified. According to the functional modelers, behaviors should always be described with their inputs and outputs, but the examples above already indicated that not everyone follows their counsel. So, if behaviors and sub-behaviors are characterized in the functional modeling way, i.e., including specific descriptions of the inputs and outputs, then we can often infer bits of categorical information from that. We already came across an example of this: The sub-behavior 'convert solar energy into electric energy' tells you that the device in question is powered by solar energy and this suggests a lot about its categorical details. But only up to a certain extent. There are no strict entailments from such dispositional descriptions to a completely specific and unique categorical description of the device. They do constrain the categorical possibilities, and sometimes rather severely so, but to actually deduce a categorical description, further background knowledge is required. Let me give

another example to clarify this point. Stone and Wood (2000) give the example of an electrical screwdriver, which has as one of its sub-behaviors 'rotate solid' where the solid in question is specified to be a screw. Strictly speaking, this description does not entail a categorical description of the screwdriver, but it is obvious that anyone with but the slightest acquaintance with screwdrivers has the background knowledge to deduce categorical facts about the shape of the screwdriver's head. To nonetheless insist in this case that there are no strict entailments is finicky. This example may be tendentious though in that the relevant background knowledge is so ubiquitous that it is hard to imagine relevant contexts in which it will not be available. That is definitely not always the case, although it is rather likely that designing engineers working in the relevant field will very often know what sort of material devices typically – or efficiently, cheaply, quickly, reliably, etc. – realize a behavior that is specified in sufficient detail. The upshot of this point is that if functional decompositions specify the inputs and outputs of sub-behaviors then this information constrains the possible categorical descriptions that may be given of the prospective artifact. It does not strictly entail specific categorical descriptions of (parts of) the artifact, but with a little extra background knowledge, which will frequently be available in a design context, it does become possible to deduce them.

(c) Besides the specifications of inputs and outputs functional decompositions can also convey categorical information in how they carve up behaviors and sub-behaviors. Specific sub-behaviors or combinations thereof are indicative of certain material devices. This only shows if the functional decomposition gets to a certain level of detail or basicality. Presumably, a functional decomposition that stays at a high level of abstraction and generality in its description of sub-behaviors does not suggest anything specific about categorical details. Decompositions at such a high level of abstractness, however, will be of little use beyond the very first stages of actual design processes, precisely because they do not get the designer any closer to a categorical description.[38] Sticking to my earlier example of computer

[38] These two claims strike me as plausible and even fairly self-evident, but I admit that I do not have compelling arguments for them, nor can I see what such arguments would look like. I can give an example of what I have in mind; one that generalizes readily as far as I can see. The behavior of a computer mouse, say 'operate a computer,' can be broken down into sub-behaviors like 'process movements of pointer,' 'process selection commands,' and 'process *Show Menu…* commands.' Such highly general and abstract sub-behaviors do not have any categorical import, but it will also be obvious that actual design processes have to go beyond these abstract descriptions to more concrete ones, like 'process click on left button' and 'process click on right button,' which do have implications about possible categorical descriptions.

mice: The higher-level behavior of different types of mice is exactly the same, but at lower levels of sub-behaviors telling differences emerge between a traditional wheel mouse, an optical mouse, and the tracking pad on a notebook computer. It is easy to multiply examples. Without going into details, it should be obvious that, at some level of detail, the sub-behaviors of a vacuum tube radio differ from those of a transistor radio, cordless phones have different sub-behaviors than ones with a cord, the sub-behaviors of diesel engines are different than those of car engines running on gasoline, let alone the sub-behaviors of fancy innovative engines running on water or sunflower oil. As before, there are no strict entailments from these dispositional descriptions to specific categorical descriptions but the categorical options are constrained. When additional background knowledge is brought into the fold it will often become possible to infer categorical information. Particular breakdowns of behaviors are typically associated with specific types of devices and engineers with the relevant experience will be able to recognize that this particular decomposition is indicative of such-and-so type(s) of artifact, component, or combination of components. These inferences do not force one particular categorical description upon a designer, because they are not deductive. It may be an explicit goal of a design process to come up with a previously unknown structure that does the same things better, cheaper, quicker, or just differently. In such cases, the conclusion that a thing with so-and-so categorical description *typically* realizes a given breakdown of a behavior is of no use, except, perhaps, in fostering inspiration for a new device. The summary of this point is thus that the specific ways in which a behavior is broken down into sub-behaviors – at least when the breakdown is given at some sufficient level of detail – constrain the possible categorical descriptions that might apply to the prospective artifact and, when supplemented with experiential background knowledge, make it possible to deduce categorical descriptions of one or more device(s) that *typically* implement the behavior or sub-behavior so decomposed.

A closing remark about (a), (b), and (c): In all three cases functional decompositions do not strictly entail categorical information (except in (a), but without further knowledge about the number and nature of the components, relative locations are an extremely non-committal form of categorical information), but they do get a designer closer to pieces of categorical information. In design contexts, the background knowledge required to deduce categorical information from functional decompositions will frequently be available. The conclusion of all this is that (IS) fails; functional decomposition cannot be done independently of any ideas about an artifact's categorical structure. (Except for when they stop at a very high level of abstraction, in which case they are of little use in actual design processes.)

This shows even clearer when we 'temporalize' the above points. To avoid

convoluted phrasings I have so far talked about how functional decompositions as finished entities imply categorical information. Once we see what this means for a *process* of functional decomposition, however, the trouble with (IS) becomes even more obvious. In a process of functional decomposition, designers (a) *choose* sub-behaviors and thereby relative locations of future components, (b) *choose to specify* behaviors with their inputs and outputs, and/or (c) *choose* to break down behaviors in specific ways. The above shows that such choices are informed by more or less specific ideas about the categorical characteristics of the artifact being designed. How else than through some notion of a prospective structure is a designer to decide that she will carve up a particular (sub-)behavior in a particular way? Below a certain level of abstractness such decisions have to be made with more or less specific ideas about categorical characteristics. At the risk of belaboring the obvious I want to add that this is not to advance the opposite suggestion that a process of functional decomposition determines *all* categorical characteristics. For each point I carefully described the sort of categorical information that is implied by functional decompositions and in no case was this information complete or fully specific.

The fourth point in this section is also the final point. I have repeatedly mentioned the importance of background knowledge for deducing structural information from singular dispositional descriptions and functional decompositions. Here I want to summarize the various sorts of background knowledge and specify them a bit more. One type of background knowledge is that which helps designers get from a singular dispositional description to one or a few possible categorical descriptions. It is experiential knowledge that particular materials, components, or objects (uniquely) satisfy a given dispositional description, i.e., can show the described behavior. A simple example is knowing that water boils at 100°C. Such knowledge makes it possible to come up with possible structures, but it is also limited in that it only takes into account one singular dispositional description. Artifacts and their parts usually have to satisfy many dispositional descriptions at once, as specified in functional decompositions. The next type of background knowledge (corresponding with (b) above) consists in recognizing the specified inputs and outputs of behaviors as being typically implemented by one or more particular (types of) components, materials, or other categorically described setups. Again this is experiential knowledge. 'Convert so-and-so alternating current to such-and-such direct current' will ring a bell for most people and certainly for electrical engineers. It is not that a categorical description of a particular type of power adaptor follows from this with necessity, but for an experienced designer it is a tiny step to infer that she might as well use such an adaptor, unless there are good reasons not to. The final type of background knowledge (see (c) above) works similar to the previous type. It is experiential knowledge that particular

breakdowns of behaviors and sub-behaviors are indicative of particular (types of) artifacts, components, or other structures. I already gave plenty of examples above. In virtue of their education and design experience designers will generally know ways in which behaviors can be and have been decomposed and implemented in particular (types) of artifacts and components. With that knowledge they will be able to generate the same or slightly different decompositions for new artifacts, which are then implicitly inspired by this categorical knowledge.

This completes my discussion of the ways in which functional decomposition depends on and determines bits of categorical descriptions of an artifact. Assumption (IS) is false; only a weaker version of it about singular dispositional and categorical descriptions can be upheld. Functional decompositions do entail categorical information about relative locations of components and processes of functional decomposition depend on more or less specific ideas about the prospective artifact's structure, at least when the behaviors are broken down to a level of sufficient concreteness and specificity, which will always be the case in practical design contexts.

### *4.4.3 Abstractness of Functional and Structural Descriptions*

Another issue from sections 4.2 and 4.3 deserves our attention: the fact that dispositional characterizations come in various degrees of abstractness (feature (E)). And, I should add immediately, so do categorical characterizations. First a few examples. A light switch may be characterized in very concrete dispositional terms by describing the exact movements of its parts in response to applied finger force, but we can also omit such details and simply say that it switches when pressed. Or, even more abstractly, turns the light on when used. Or, given appropriate contexts, that it alerts a burglar, prevents a theft, or saves the family's precious heirlooms (using a Davidsonian example). Each description abstracts away from more specifics of the light switch and becomes more multiply realizable. Nonetheless, all of these things describe (possible) behavioral roles of a light switch. Now for categorical characterizations. By way of a far-fetched theoretical possibility, a maximally concrete categorical description of an artifact would mention all of its constituting atoms and their relative locations. More practicable categorical descriptions just contain the kinds of materials the object is made of and its spatial dimensions. Or, even more abstractly, describe it as, say, a cubic solid object. Abstract categorical descriptions are also multiply realizable in the sense that they fit many different objects.

It is interesting to note that when we try to go even more abstractly, categorical descriptions automatically start to get a dispositional flavor. Door, cup, desk, car, dice, and other artifact kind terms may not be names of

functions *expressis verbis* like screwdriver, computer, or percolator, but they are intimately associated with functions. Depending on one's favored analysis of artifact kind concepts, they may even analytically entail them.[39] One might think of artifact kind names as categorical descriptions, precisely because they name *kinds* of things, which typically gives a categorical description. But since artifact kinds seem so closely related to technical functions, it is also plausible to think of kind names as dispositional descriptions. I will not choose one or the other option here, since the choice is inconsequential for what I want to say. This observation does suggest that there may be a maximum level of abstractness for categorical descriptions of artifacts, above which descriptions become (implicitly) dispositional. It would be a mistake, however, to think that only descriptions of properties above this level are dispositional while everything below it automatically classifies as categorical. The upper level of categorical abstractness does not coincide with the lower level of dispositional abstractness.[40] To see this, one only has to realize that the engineering sciences, as well as biology, chemistry, and physics, even fundamental physics, are full of dispositional descriptions of properties. Properties such as conductivity, elasticity, and rigidity are all clearly dispositional and many philosophers have argued that even mass, charge, and spin are best understood as being dispositional characterizations. In fact, some have claimed that the notions we use to refer to fundamental particles are essentially dispositional. A recent example is Mumford (2006), who argues that the dispositions of quarks and electrons are ungrounded, i.e., cannot be accounted for in terms of categorical descriptions of their constituents.

Dispositional and categorical descriptions are thus not latched onto fixed levels of nature or levels in the hierarchy of the sciences; they are at home at every level. Dispositional descriptions are not necessarily more abstract than categorical ones. Lycan (1999), crediting Dennett (1978), elaborates this point by suggesting that levels of nature are continuous and that the distinction between dispositional and categorical is not absolute but relative to levels of nature. He uses levels of nature in a loose sense as standing for levels of abstractness in our idioms for describing nature. Says Lycan: 'The 'function'/'structure' distinction *goes relative*: something is a role as opposed to an occupant, a functional state as opposed to a realizer, or vice versa, only *modulo* a designated level of nature' (ibid.: 50, where ''function'/'structure' distinction' should be read as

[39] The problem of artifact kinds is more difficult than it may seem at first sight. Some of my thinking about it is in the concluding chapter of this thesis.

[40] This talk of levels of abstractness is to a certain extent a figure of speech. I do not have a measuring stick for abstractness, but I do think we can sometimes pass judgments on the relative abstractness of two descriptions. I use that ability in my examples.

dispositional-categorical distinction). Given that I have followed Mumford in defending a sharp distinction between dispositional and categorical descriptions, we should not interpret Lycan as saying that the same description can be both dispositional and categorical relative to different levels. What he is emphasizing is that a property can both be described dispositionally relative to a higher level of nature, i.e., in more abstract terms, and categorically relative to a lower level of nature, i.e., in more concrete terms. The same property would come out as dispositional when described more abstractly as 'conductivity' and as categorical when described more concretely as 'having *n* free electrons.' So the moral is that the dispositional-categorical distinction is not absolute but relative. It does not have a fixed cut-off point somewhere near, say, where chemistry or physics starts. Dispositional and categorical descriptions both come in various degrees of abstractness and can be used at all levels of science or nature.

What does this mean for functional decompositions? First, since functional decompositions are made up out of ordered singular dispositional descriptions, they, too, can come in various degrees of abstractness. More precisely, the dispositional descriptions in them may vary in abstractness. This ties in nicely with the suggestion I made earlier, that one of the things functional decompositions do is reduce the amount of abstractness of dispositional descriptions. When the initial description of an artifact's intended behavior is very abstract, a functional decomposition reduces this abstractness by describing sub-behaviors more and more concretely. We will come back to this suggestion later when we look into the various relations between a behavior and its sub-behaviors. This feature of functional decompositions is important because more concrete descriptions constrain the possible categorical descriptions that may apply to a prospective artifact or component more strongly. Saying that a device should make a ringing sound when pressed allows for significantly less categorical descriptions than merely saying that it should alert someone when used. By making dispositional descriptions more concrete and specific, more and more categorical possibilities are excluded and that makes it easier to come up with one final categorical description. This, of course, contributes directly to the accomplishment of a design task.

A final consequence of the relativity of the dispositional-categorical distinction that I want to stress is that we should jettison any hierarchical intuitions we might have about dispositional descriptions always being at a higher level than categorical ones. Such intuitions are mistaken; everything about the distinction goes relative. We sometimes account for an object's dispositionally described properties in terms of lower-level categorical descriptions of properties, but we may as well explain dispositions in terms of lower-level dispositions had by the artifact's components. At other times we do

it the other way around. A molecular structure can be accounted for in terms of dispositional properties of atoms. There is no simple hierarchical division between dispositional and categorical talk; both occur at multiple levels of abstractness and explanatory relations between them run in both directions.

### *4.4.4 When Functional Decomposition Stops*

Anyone who has kept count will know that there is still one feature left from sections 4.2 and 4.3: the one I labeled (F), which was that functional decompositions stop at some point when the identified sub-behaviors are considered to be sufficiently basic or atomic. Frankly, I have little more to say about this feature than what I already said above, except that we can now see how that dovetails with the more detailed analysis in the present section. I will briefly repeat what I said above and then add one more comment. The stopping point for a process of functional decomposition is relative to engineering disciplines and to the design task at hand. Behaviors that count as basic for one discipline need not do so for another. And where most design assignments can stop when the lowest level sub-behaviors coincides with the behaviors of standardized off-the-shelf component, other innovative projects will want to decompose further. We can now see that this procedure is a specific instance of phenomena (b) and (c) I described in section 4.4.2. If a designer recognizes a sub-behavior and its input and output as one that is typically taken care of by some known component or material, she may as well decide to use one of these, especially if they are off-the-shelf, standardized items that fellow designers tend to always use in similar design situations. Or if she recognizes a sub-behavior as one that is liable to such-and-so default decomposition, she might as well go for this default option and use the component that is usually associated with it. Functional decomposition, then, stops when the designer has an idea about what sort of things and materials she can use to realize (most of) the lowest-level sub-behaviors.

This may sound as if the entire design transition from an abstract dispositional to a concrete categorical description is already complete when functional decomposition is over, but that is not at all the case. Here are a few things that still need to be sorted out: For some sub-behaviors it may still be unclear how they can be implemented. What is more, knowing one or a few potential components for each individual basic sub-behavior in no way adds up to a satisfactory description of an artifact. Decisions have to be made about which of the various categorical possibilities for each sub-behavior will be chosen, it has to be sorted out which combinations of sub-behaviors can be realized by which components, and, most importantly, these decisions have to be made so that the resulting components and combinations are physically feasible and compatible, do not adversely interfere with one another, are

affordable, sufficiently efficient, reliable, cheap, fast, safe, silent, noisy, small, big, light, heavy, and whatnot. In other words, there is still plenty of work to do in a design process when the functional decomposing is over. And this is true regardless of the fact that the designer has advanced significantly to a categorical description during the process of functional decomposition.

### *4.4.5 Concluding Remarks*

In sum: If you are willing to accept my analysis of the distinction between dispositional and categorical descriptions (in Mumford fashion) you can account for a lot of the features of functional decomposition we discovered in our review of examples and more theoretical treatments of the topic in the engineering literature. That was the argument of this section. I showed how functional decomposition cannot be done completely independent of ideas about the categorical structure of a prospective artifact, *pace* design methodologists who seem to endorse (IS). Categorical information necessarily seeps into a functional decomposition. The relativity of the dispositional-categorical distinction tied in nicely with the fact that dispositional descriptions come in various degrees of abstractness. My earlier suggestion that what counts as a sufficiently basic sub-behavior is relative to engineering discipline and design task also turned out to sit well with the dispositional-categorical analysis.

All of this is fine and it has provided us with telling clues as to why functional decomposition would be an important part of a successful design process, but we can do even better. Up until now we have relied on our intuitive grasp of the notions of behavior, sub-behavior, and the relations that may hold between them. While this has been far from unrewarding it is now time to unpack those notions. In the next section we take a brief detour to establish what sort of things behaviors are, ontologically speaking. This information is required because not all things can stand in the same sort of relations; propositions can entail one another but people cannot. When we know what sort of things behaviors are we can map out different possible relations between behaviors and sub-behaviors. An inventory of these possible relations yields a variety of possible meanings of functional decompositions that have never before been distinguished in the literature.

## 4.5 Behaviors and Events

To further develop our thinking about functional decomposition we should find out what sort of relations can hold between behaviors and their sub-behaviors. In order to achieve that aim we need to get clear on is the nature of behaviors; where, in the grand ontological scheme of things, do they fit in?

Behaviors are not typically thought of as a distinct ontological category besides, say, material objects, abstract objects, properties, places, and times.

They are subsumed under the category of events, which does constitute a somewhat reputable ontological category, notwithstanding various attempts to reduce them to times, places, or properties. Philosophers hardly agree on how best to construe events, as testified by the multitude of opinions collected in Casati and Varzi (1996), but I do not want to enter into debates on the exact nature of events. Instead I will help myself to a few basic and hopefully not overly controversial assumptions about behaviors and events. First, I will assume that there are complex events, i.e., events composed of more than one proper part. An example could be someone's writing a paper, which has as its parts the writings of sections, sentences or even words, all of which may take place at different and unconnected times and places. Second, events must be individuated fairly narrowly to deal with the problematic cases that coarse-grained individuation runs into, such as those discussed in Goldman (1970: 2–10). For instance, think of John's saying 'hello' loudly. On a coarse-grained individuation of events, this would surely be the same event as his saying 'hello.' The problem with this suggestion is that these two events may well have different causes and effects. Suppose John's being upset causes his saying 'hello' *loudly* while the fact that he just spotted his neigbor causes his saying 'hello'. Furthermore, his saying 'hello' *loudly* scares away a bird in his garden while his saying 'hello' causes his neighbor to wave his hand. Since the causes and effects of his saying 'hello' loudly are so different from those of his saying 'hello,' it is hard to maintain that these two events are identical. Narrow individuation makes it possible to identify John's saying 'hello' as a different event than his saying 'hello' loudly. It is equally obvious, however, that the two events are intimately related and therefore an additional story must be told as to what this relation consists in. More on that later. My third assumption is that there is no problem about different events occupying the exact same region of space and time. Most microwave ovens rotate food while heating it. The food's rotating and its being heated are two distinct events occurring in the exact same stretch of time and region of space. (John's saying hello 'loudly' and his saying 'hello' is another example.)

This will suffice for the present purposes. We are now ready to launch into the next main task of the chapter, which is to disentangle different relations that may hold between a behavior and its sub-behaviors. These different relations give rise to different senses of functional decompositions.

### 4.6 The Many Senses of Functional Decomposition

Before we start I need to get two possible misunderstandings out of the way. First, I want to make it explicit that I use 'decomposition' in a broad sense, as an umbrella term for the various meanings I am about to explore. This may lead to confusion if you think about composition and decomposition as

having to do exclusively with mereological (part-whole) relations. My reason for stretching the meaning of the term is that that is exactly what engineers do, although presumably not always consciously so. By looking carefully at the examples of functional decomposition I gave above, you will see that they do not all fit the mold of a purely mereological analysis. Taking engineers' use of the term 'functional decomposition' seriously forces one to broaden its meaning. (Alternatively, one may accuse designers and design methodologists of being sloppy about their language and introduce a new term for each of the senses of functional decomposition I am about to distinguish. I have no quarrel with that strategy in principle; it just strikes me as unnecessarily rude towards designers and methodologists. The important point for me is not terminology, but the underlying differences between the various things that engineers tend to group under the heading of functional decomposition.)

The second misunderstanding has to do with a sense of functional decomposition that I will not attend to here. It is what I would call physical decomposition from a functional perspective. Many artifacts have numerous physical parts that take care of the various parts and aspects of their correct functioning. Depending on the functional perspective you are interested in, you may view different collections of physical parts as composing a particular physical subsystem. Varying perspectives will give rise to different ways of carving up the artifact in subsystems and the same physical parts may be part of more than one subsystem. A car can serve as an example. Modern cars may be thought of as having, e.g., a propulsion system, brake system, steering system, and lighting system. The car's tires will come out as part of the propulsion system, the brake system, and the steering system. It seems to me that this is a perfectly legitimate sense of functional decomposition, but the misunderstanding is that some people think that it is the *only* possible sense. So we find McLaughlin (2006: 99) asserting, without further argument, that 'functional analysis is the analysis of a structure from a functional perspective: the parts of a system have the function of contributing to the relevant behavior or capacity of the system. These function bearers are parts of the containing system, but their functions are not in any sense parts of its function.' The work on functional decomposition I discussed above makes it clear that engineers differ. Physical decomposition from a functional perspective is not what they have in mind when they talk about functional decomposition. As far as I can see they are quite right, for once you accept that functions are (intended) behaviors, that behaviors are events, and that events can stand in many different relations to one another, it should be evident that there are more literal senses of functional decomposition as well.

With these misunderstandings cleared up, let us look at how behaviors and sub-behaviors may be related. The first way is for them to be related as a set and

members of this set. An artifact's overall 'behavior' may really be a collection of behaviors, which are not clearly related in any sense other than they are performed by the same artifact or component.[41] Think of a multifunctional artifact. Even though such an artifact literally has more than one function, we may still loosely speak of *the* overall behavior of that item, which would then in fact refer to a number of behaviors. A fancy KitchenAid may serve as an example. If we speak of the function of such a device, we are in fact referring to a collection of behaviors it can exhibit, but these different behaviors are not in any sense part of, caused by, or identical to each other. A functional decomposition of a KitchenAid may nonetheless start with some catchall term for its overall function and then subsume the various sub-behaviors it is supposed to perform under that.

This relation is not terribly interesting from an ontological perspective, since set membership is so loose a relation and very easy to come by. Basically, all that is required for it is that somebody defines or thinks about these behaviors as somehow belonging together. This is not to say that functional decomposition in this sense is just about conjuring up arbitrary collections of behaviors. In design processes one will need to have good objective reasons for grouping together an otherwise unrelated collection of behaviors as being sub-behaviors of some artifact's overall behavior. One such reason may of course be the fact that a client desires a multifunctional device such as a KitchenAid, but in further stages of the design process there may also be a need for multifunctional components and a decomposition of their overall behavior would also involve the set-member relation. Other reasons for thinking about different behaviors as somehow belonging together may be their spatial or temporal conjunction or contiguity. Such reasons are by no means sufficient conditions, since many coincident or adjacent events are completely unrelated and there may not be any reason to think about them as members of the same set (although one could always stipulate a set with such events as its members, if one wanted to). What all these reasons have in common is that they group behaviors together by their extrinsic properties, e.g., their being performed by the same physical object, contributing to some goal one has in mind, or standing in specific spatiotemporal relations. This is the reason why I have been saying that this form of functional decomposition is not terribly interesting. Without further knowledge of, e.g., how the behavior is supposed to be performed, the clients' needs, or similar artifacts that have performed the same behavior(s), the

[41] Depending on your opinions about event individuation and composition, the very fact that the same artifact or component performs these different behaviors may be a reason to see these behaviors as composing one bigger complex behavior. Since I do not want commit myself to any particular criteria for event individuation or composition, I will leave this issue open.

behavior itself does not give a designer much to go on for coming up with sub-behaviors in this set-membership sense (unless, of course, the initial description of the behavior is already a description of a collection of disjunct behaviors). Because this sense of functional decomposition is so open-ended, it might do a lot of different things. Identifying the members of a set might be a way of specifying an overall behavior further (feature (A) in section 4.2), through it categorical information might seep into a process of functional decomposition (features (B) through (D)), and it might reduce the abstractness of a given description of behavior if the members of the set are described more concretely (feature (E)).

The second relation between events that we will consider gives rise to the most straightforward sense of functional decomposition: the mereological part-whole relation between events. If behaviors can be complex, they can have proper parts and these parts then stand in the parthood-relation to the behavior of which they are parts. And this may continue down through several levels of nested parthood. My writing this chapter has its proper parts my writing the first section, the second section, and so on. And my writing the first section has as its proper parts my writing the various paragraphs of the first section. But perhaps my writing this chapter has as its proper parts my writing the first draft, revising that into a second draft, and so forth. Or perhaps we should say that it has as its proper parts various periods of time in which I write chunks of this chapter. As for artifact-related examples, a car's assembly has as its parts the assemblies of its various physical parts. Or a thermostat's behavior has as its parts measuring of the temperature, comparing it to a norm, deciding whether the temperature deviates from the norm, and, depending on the outcome, switching a heating or cooling system on or off. A coffee maker's behavior would have parts like the heating of water, percolating, and collecting the coffee. Functional decomposition in this sense literally means finding the parts, or sub-behaviors, of the overall behavior.

Events can have both spatial and temporal parts. The two types of parts are independent. An event's temporal parts can occur at the same place and its spatial parts can occur at the same time. The examples above exemplify this. How event-parts are to be individuated is unclear. If we assume that all events take place in space and during some length of time[42], it is always possible to slice up an event into temporal or spatial parts, but it is far from obvious that such divisions will cut an event at its joints, so to speak. If not nonsensical

[42] This assumption is not obviously true. Does a machine's breakdown take place exactly where the dysfunctional component sits or does it fill the entire space the machine occupies? Does a custom built car's production take place at the same location as that car's failing to meet the customer's

it is definitely weird to talk about the right and left half of, say, an electrical current's running through a copper wire. To the extent that it does make sense to identify spatial parts of events, their individuation seems to be parasitic on the individuation of the objects participating in them. The spatial parts of a clock's telling the time may be the movings of its big and small hand. For temporal parts, the situation is different. If an event takes place during non-contiguous intervals of time or if it involves quantitatively different sorts of happenings, we are inclined to identify these episodes as its temporal parts. For instance, a pulse generator's behavior is interrupted by intervals of silence and we may identify each pulse generation as a temporal part of its overall behavior. A piston's movement could have as parts its moving up its moving down, because an upward movement is something quantitatively different from a downward movement. But the individuation of temporal and spatial event-parts is not our topic. It suffices if the distinction between these two types of event parts is reasonably clear; we can leave precise individuation criteria to the metaphysicians.

In addition to temporal and spatial parts, some events seem to have parts that occur at exactly the same places and times. Earlier I gave the example of a microwave oven that rotates and heats food. The same happens with ball bearings; they, too, rotate and warm up. Windows both let through light and block wind. A printer both leaves marks on a sheet of paper and transports the sheet from the paper feed to the output. *If* we want to say that such behaviors compose one complex behavior, then its parts are individuated by their being qualitatively different sorts of changes. Changes in temperature and changes in spatial properties, changes in velocity and changes in chemical structure are qualitatively different sorts of things even though they may occur at the very same time and place. So we end up with three sorts of event-parts: temporal, spatial, and qualitative.

Straightforward as the idea of events having parts may sound, it does not automatically translate into a crisp procedure for successful functional decomposition. The major obstacle is individuation and composition criteria for events: How are behaviors and their parts to be individuated and when do two or more individual behaviors make up a new individual complex behavior? As long as we do not have answers to these questions – which, for all I know, is our predicament – we cannot draw the line between the first sense of functional decomposition and the second. On the one hand, if composition conditions are really strict we end up with an event ontology that is sparsely populated with

expectations? Is the latter event perhaps momentary, in the sense of lacking temporal extension? Can events have vague spatiotemporal boundaries? Those are difficult questions, but they are beyond the scope of this chapter and must remain unanswered here.

simple basic events. A lot of our colloquial talk about complex events would then be strictly speaking incorrect. For example, my suggestion that a car's assembly is one big complex event would be mistaken because the assembly is merely a set of simpler events. The same holds for the writing of a paper. That, too, would not be one event but a mere collection of simpler events. On the other hand, more liberal composition criteria result in a densely populated ontology that renders talk about complex events literally true. And, of course, there are all manner of options in between the very strict and the very liberal. I do not know which of these options is more preferable.

Instead of trying to provide a compelling account of behavior individuation and composition, let me offer a list of features of events that we seem to use, though not without exception, as grounds to group behaviors or events together. None of these features are necessary or sufficient conditions for behavior composition, but they are relevant considerations. Depending on one's taste in ontology, one could see these features as genuine composition conditions or as merely subjective grounds for taking behaviors to be members of some set. The list is not supposed to be exhaustive, but it covers the more conspicuous and important features we seem to use to group behaviors together.

(1) Being performed by the same object, component(s) of an object, or an otherwise related collection of objects.
(2) Being performed in the same spatiotemporal region.
(3) Being performed in overlapping or adjacent spatial regions.
(4) Being performed at overlapping or consecutive times.
(5) Contributing to the same effect or intended goal.

Using these five features, decomposition of a given behavior could thus mean: (1) finding simpler behaviors that various components of an artifact or artifacts in a system of artifacts can perform, (2) distinguishing the qualitatively different changes involved in the overall behavior, (3) disentangling its spatial parts, (4) breaking it down into temporal parts or phases, or (5) coming up with simpler behaviors that bring about changes that accumulate to a given effect or desired goal.

All of these things can be ways to specify a behavior in more detail (feature (A) from section 4.2). They may also import bits of categorical information if the decomposition steps are inspired by and depend on more or less specific ideas about the structural details of the prospective artifact and its components (features (B) through (D)). A brief example to clarify this point: if spatial and temporal event parts are indeed parasitic on physical parts, a designer will need to have some notion of the parts that the artifact she is designing is going to

have in order to identify spatial and temporal parts of an overall behavior.

To introduce the third sense of functional decomposition we have to work our way through some event metaphysics first. That takes up the next section; in section 4.8 I get back to the main line of argument and spell out the third sense of functional decomposition.

### 4.7 Intermezzo: Level Generation and Event Supervenience

In the philosophy of action, it has long been recognized that some actions and events stand in a by-relation. People often perform some action *by* performing another. Davidson's famous example is a case in which someone alerts a burglar to the fact that she is home by illuminating a room, which she does by turning on the light, which she does by flipping the light switch (Davidson 1980, Essay 1). Or, for a more violent example, someone's killing the victim by shooting him by pulling the trigger by moving her finger. The by-relation is equally important for understanding the possible relations between artifact behaviors. Some computer mice translate the movements of your hands to an electrical signal by sensing reflected light; others do it by tracking the movements of a rubber ball. Radios used to amplify sound by whatever it is vacuum tubes do; later they did it by what transistors do. In fact, I think the by-relation is key to understanding functional decomposition better, since one of its tasks is getting from abstract functional talk to concrete functional talk and that is what the by-relation does. More abstract events are done by less abstract events. Choosing the lower-level behaviors by which higher-level behaviors are realized seems like exactly the sort of thing that a designer has to do.

Always keen on distinctions, philosophers have identified several subtly different forms of the by-relation. Readers who are put off by technical definitions are advised to skip the rest of this section and continue reading at the start of the next section. The points to remember are that the by-relation can be analyzed in terms of *event supervenience*. A higher-level event that is performed *by* a lower-level event is said to event-supervene on that lower-level event. There are different forms of event supervenience.

Alvin Goldman (1970) proposes to analyze the by-relation in terms of what he dubs *level generation*. One action can level-generate other actions and these higher-level actions are then said to be performed by the lower-level ones. Together, the following three items are a necessary condition for events to stand in a relation of level generation to each other. (1) The events must not be subsequent to one another; they must take place in the same interval of time. (2) They must not be completely independent of each other, i.e., it is incorrect so say of them that someone did one *while also* doing the other. (3) They are not proper parts of each other. Goldman distinguishes different types of level generation.

Building on his work, Peter Menzies (1988) shuffles around and adds to Goldman's typology of level generation and he generalizes the notion to apply to events in general, instead of just actions. He talks about *event supervenience*, which is more or less the mirror image of level generation. Where Goldman says the more basic event level-generates the less basic event, Menzies says the less basic event supervenes on the more basic one. I will not be concerned with details and internal criticism of either of their lists but simply compose my own list of types of event supervenience, borrowing freely from theirs and adopting Menzies's supervenience terminology. I think the following forms are relevant in the present context; we will have opportunity to see why in the next section.

(1) *Causal supervenience*: Behavior $b^*$ causally supervenes on behavior $b$ iff (a) $b$ causes effect $e$, and (b) $b^*$ consists in $e$ being caused. For example, the unlocking of a door causally supervenes on the turning of a key. The turning on of a television causally supervenes on the pushing of a button.
(2) *Conventional supervenience*: Behavior $b^*$ conventionally supervenes on behavior $b$ iff the performing of $b$ in circumstances $C$ (possibly null), together with a rule saying that $b$ done in $C$ counts as $b^*$, guarantees the performing of $b^*$. A traffic light's signaling 'stop' conventionally supervenes on its turning red. A bridge's being safe conventionally supervenes on its being able to carry some definite load, assuming there are safety standards specifying when bridges are safe.
(3) *Simple supervenience*: Behavior $b^*$ simply supervenes on behavior $b$ iff the performing of $b$ in circumstances $C$ guarantees the performing of $b^*$. An alarm's going off simply supervenes on its making high-volume sounds that are audible to the human ear.
(4) *Logical supervenience*: Behavior $b^*$ logically supervenes on behavior $b$ iff $b$ logically entails $b^*$ but $b^*$ does not logically entail $b$. A cogwheel's rotating at some speed logically supervenes on its rotating at 60 rpm.
(5) *Typical causal role supervenience*: behavior $b^*$ supervenes in terms of typical causal role on behavior $b$ iff (a) $b$ involves scientifically more fundamental properties or entities than $b^*$ and (b) the properties or entities involved in $b$ fulfill a typical causal role such that a statement that they occupy this causal role entails $b^*$. A wire's conducting electricity supervenes on the behaviors of its electrons because these behaviors underlie manifestations of the wire's conductivity.
(6) *Mereological supervenience*: Behavior $b^*$ mereologically supervenes on a complex event $b$ (consisting of $b_1, b_2, \ldots, b_n$) iff (a) the objects involved in $b$ are the physical constituents of the object involved in $b^*$ and (b) the statement that event $b$ occurs, together with the appropriate bridge laws, entails that event $b^*$ occurs. The raising of a body of gas's temperature

mereologically supervenes on the increasing of the average kinetic energy of its constituent molecules because temperature is average kinetic energy.

A few brief remarks about this list. First and most importantly, I do not want to claim any sort of completeness or finality for this list. The items on it seem to me to be *prima facie* different sorts of event supervenience, which are useful for thinking about the relations between artifact behaviors and sub-behaviors. Second, the items on my list are not mutually exclusive. In particular, (1) and (2) are special cases of (3), in which the circumstances include (1) the holding of a particular causal relation and (2) the existence of a particular (social) convention. Third, instances of event supervenience may be combined to form a tree of supervening behaviors. An alarm going off simply supervenes on the producing of sounds, which logically supervenes on the producing of the specific sounds that are produced, which causally supervenes on someone's activating the alarm, which causally supervenes on someone's hitting a switch. Fourth, it is not entirely obvious that logical supervenience truly belongs on this list, for it seems to be the only relation that does not straightforwardly translate into by-locutions. At best, it is idiosyncratic to say that a car moves at some speed by moving at 50 km/h.

With these qualifications in mind, let us look at how we can use the items on this list to distinguish yet more different senses of functional decomposition.

### 4.8 More Senses of Functional Decomposition

We can now continue the list of different senses of decomposition that was interrupted in the previous section. The general idea of the third sense of functional decomposition is that functional decomposition means finding supervenience bases for some behavior. (In Goldman's vernacular, finding lower-level behaviors that would level-generate the desired higher-level behavior we are decomposing.) Each form of supervenience on my list gives rise to a slightly different form of functional decomposition. This general idea is pretty straightforward. Since the previous section already contained quite a few examples I can be brief here. Someone with more hands-on experience at functional decomposition will surely be able to multiply examples. Let's take each type of event supervenience in turn.

First, finding a *causal supervenience* basis requires for a behavior or sub-behavior to be described as the bringing about of some effect or goal. Decomposing such a behavior would then mean coming up with another behavior that would bring about this effect. Roughly, it is finding a means to an end, thinking of a particular way to bring about an effect. Suppose we are decomposing the sub-function 'closing an electrical circuit,' which may itself be part of a causal supervenience base for, say, the ringing of a doorbell.

One straightforward way to do this is to position a piece of metal between the two terminating points of the circuit. This positioning would be a causal supervenience base behavior because the positioning of the metal causes the circuit to be closed. Here are some other examples: Running electricity trough a heating element causes the element to warm up, which may cause water in a kettle to start boiling if the heating element is inside the kettle. Pushing a button on a remote control causes the remote control to transmit an infrared signal to a stereo.

This form of functional decomposition takes designers from more abstract descriptions of behaviors to more concrete ones. That is to say, it provides one of the mechanisms by which we can account for feature (E) from section 4.2. And since the identification of possible causal supervenience bases will typically require background knowledge about possible structures of the sorts described in section 4.4.2, my points about features (B) through (D) and the role of background knowledge in functional decomposition are also confirmed.

Second is *conventional supervenience*. Some artifact behaviors count as higher-level behaviors because of official or quasi-official norms and standards that an artifact must satisfy. Think of safety regulations, performance characteristics, or environmental criteria. Functional decomposition in this sense boils down to finding specific behaviors that will adhere to such norms, standards, and criteria. In the Netherlands there is an energy use classification scheme for household appliances with categories ranging from *A* for very efficient, sustainable use of energy to *F* for inefficient, unsustainable use of energy. If one sub-behavior of a washing machine would be to use energy sustainably, that behavior would conventionally supervene on its using energy according to the class *A* specifications. Like the previous sense of functional decomposition this one, too, reduces the abstractness of dispositional descriptions. Designers also need background knowledge about what sort of structures will be able to meet the given norms, standards, or criteria. With that knowledge they can identify possible sub-behaviors for the behavior they are decomposing.

Third, *simple supervenience* involves finding behaviors that, in relevant circumstances, will realize the desired behavior or sub-behavior. To distinguish simple supervenience from causal and conventional supervenience, the circumstances in question should neither be the holding of a certain causal relation between the base behavior and some effect the being brought about of which is the supervening behavior, nor the existence of some rule saying the base behavior counts as the supervening behavior. That still leaves open many possibilities. Let me mention a few. A nail clipper converts the force of your finger pressing to a larger force that cuts off the nail. This behavior is realized through the behaviors of the nail clipper's handle and the way in which it is attached to the blades. This configuration realizes a lever effect. So we have

the supervening behavior of force conversion being realized by what the nail clipper's parts do. Or think of a car's acceleration. The car's ability to accelerate from 0 km/h to 100 km/h in 8.2 seconds could be a supervenience base for its outperforming the competitor's latest model. In the same way as the previous two, this sense of functional decomposition reduces abstractness and requires some amount of background knowledge. In some cases, like in the example of the car's acceleration I just gave, it might also just be a matter of further specification of a function. So this sense of functional decomposition accounts for feature (A) from 4.2 as well.

Fourth, *logical supervenience*. This is perhaps the least interesting sense of functional decomposition because it only involves stipulation or precisification. The coffeemaker's heating the water logically supervenes on its heating the water to around 80°C and the packing machine's combining the tiles into lots supervenes on its combining them into lots of ten tiles. It is not unimportant to specify the exact values of variables and quantities, but this activity does not give us particularly deep insights in how engineers get through a design process successfully. In this form, functional decomposition is largely about making descriptions of behavior more precise and in so doing it is related to feature (A) from 4.2 and to the reducing of abstractness (feature (E)).

Fifth, *typical causal role supervenience* is the way to get from properties at higher scientific levels to more fundamental properties. This would be the type of functional decomposition that enables designers to get from a dispositional description of specific properties to a categorical description in terms of the types of material or molecular structures that have these dispositional properties. When they are looking for something with so-and-so conductivity, elasticity, stiffness, resistance, etc., finding a material with these dispositional properties would be decomposition in terms of typical causal role.

Although Menzies is primarily concerned with the relation between dispositional properties and their categorical bases, a natural extension for this type of supervenience is to let it also cover cases where designers jump from sub-behaviors that are considered to be sufficiently basic to a standard 'off-the-shelf' component that is known to be able to exhibit these behaviors or simply is the default way of doing things (cf. feature (F) and section 4.4.4 above). By way of examples, think of screws, bolts, transistors, buttons, etc. It may be perfectly possible to invent new items that can perform the behaviors of these items, but normally there is no need whatsoever to do so. This fifth form of supervenience thus gives rise to a fairly specific and narrow sense of functional decomposition that has its natural place near the end of a process of functional decomposition. Once the bigger picture of how an artifact is going to carry out its desired behavior is in place and the decomposition is at the level of specific behavior manifestations, this form of functional decomposition takes the final step to

a categorical description of the required components and materials. That does not necessarily make this step particularly easy or straightforward, since it may well be the case that the disposition manifestations described in the functional decomposition are very hard or even impossible to realize with any (known) components, materials, or combinations thereof.

Because this sense of functional decomposition in effect takes the step from a dispositional description to a categorical one, engineers and design methodologists would probably not call it functional decomposition. After all, we saw that they are keen on keeping functional decomposition free from structural information (even if that turned out to be harder than they seemed to realize). So perhaps we should conclude that this relation between behaviors does not translate into a sense of functional decomposition in the way engineers use the notion. In that case, consider the previous paragraph a bonus comment that offers some insight on what is involved in taking the step from basic sub-behaviors to categorical descriptions of materials or components that can realize those sub-behaviors *after* the process of functional decomposition has finished.

Sixth and finally, *mereological supervenience* introduces a form of functional decomposition in which you decompose a complex behavior of an object into behaviors of that object's physical components. Menzies takes this type of supervenience to apply primarily to cases where a higher-level behavior of a complex system composed of a great many of identical 'atomic' constituent objects is accounted for by the lower-level behaviors of these atoms. Hence his mention of bridge laws and the examples of the temperature of a body of gas and the behavior of an economy. This type of supervenience is hardly relevant here. The artifacts we are considering are not complex higher-level systems made up out of a great many roughly identical atomic parts. Some artifacts may contain or use gases or fluids that can be so analyzed, but that is irrelevant in a design context since designers do not (yet?) design or manipulate individual atoms or molecules. Changes in the temperature of a fluid may supervene mereologically on changes in kinetic energy of its parts, but a decomposition that describes this relation is of little use for engineering design because it is at too low a level of detail.

It seems to me, however, that nothing prohibits this type of supervenience from also applying to behaviors of physically complex objects with various parts and components, where the object's components perform lower-level or simpler behaviors. The idea of bridge laws would then have to be interpreted rather loosely, for we do not normally think about the behavior of a pair of scissors and those of its two blades as connected by a bridge law. But once we accept this idiosyncratic usage this sense of functional decomposition becomes very similar to the straightforward mereological sense of functional decomposition

described in section 4.6, in which decomposition simply meant identifying the parts of a behavior. There is one significant difference though. The mereological supervenience sense we are looking at here identifies behavioral parts *at a lower level of abstraction*, whereas the earlier mereological sense carried no such implications about levels. The exact difference between the two senses may not always be easy to tell in practice, especially since the notion of levels of nature or levels of abstraction is imprecise. I trust, however, that the reader will grasp the intuitive difference between the two. This sense of functional decomposition, then, takes care of several things at the same time. It may be used to specify an overall behavior further (feature (A) from 4.2), to break the overall behavior down into simpler parts, and to move to more concrete descriptions of behavior (feature (E)).

## 4.9 What Good Is Functional Decomposition?

This has been a lengthy chapter and its overall argument stretches out over several sections. It will be helpful to pull everything I have said together and look at how it combines into a coherent account of the role of functional decomposition in engineering design processes. First a brief recap of the reasoning that led to the project of this chapter and then the picture that emerges from this chapter.

The basic problem that has to be solved in a design process, you will recall from chapter 2, is to get from an initial description of an artifact in abstract and largely dispositional terms to a concrete and categorical description, or at least a description that entails a complete categorical description in the given context. Design methodologists counsel that the method of functional decomposition be used to make (part of) this transition. The question for this chapter was how functional decomposition accomplishes that task. To answer that question we took the following steps. In sections 4.2 and 4.3 we reviewed what design methodologists say about functional decomposition. This led to a list of observed features of functional decomposition that have to be accounted for. In section 4.4 I showed how the distinction between dispositional and categorical descriptions of artifacts makes sense of a number of the observed features. Specifically, use of the distinction enabled us to see that functional decomposition cannot be done independently of ideas about the categorical make-up of the artifact being designed and that the varying abstractness of dispositional descriptions ties in nicely with the relativity of the dispositional-categorical distinction. In sections 4.5 through 4.8 we looked at the 'mechanism' of functional decomposition. Since the point of functional decomposition is to identify sub-behaviors that will realize a given behavior, I sought out the types of relations that may hold between behaviors. These different types of relations gave rise to different senses of functional

decomposition, all contributing to some aspect of the transition from an initial abstract dispositional description to a concrete categorical description of an artifact.

The result of this is a fairly comprehensive account of functional decomposition in engineering design. If this account is correct and if design methodologists have been right in their endorsement of functional decomposition, it should now be possible to show how functional decomposition contributes to bridging the gap between a design assignment that contains mostly dispositional requirements on the one hand and a categorical blueprint for a particular artifact on the other. Throughout the chapter I have already commented on this quite a bit, but I will draw the important points together here.

First a brief and general answer and then elaborations on a few specific aspects of that answer. The general answer is this. The different forms of functional decomposition (1) specify the overall behavior(s) of the artifact being designed and the constraints on its behaviors, (2) break the overall behavior down into more basic parts (or members of the overall set), and (3) translate abstractly formulated behaviors into more concrete descriptions of sub-behaviors. The process continues until the sub-behaviors at the lowest level are considered to be sufficiently basic or – if there is such a thing – truly atomic. The resulting functional decomposition is a hierarchically ordered collection of descriptions of behaviors that the artifact's components, whatever they will be, must be able to carry out. This endpoint is considerably closer to a categorical description of an artifact than the desired behavior with which the process of functional decomposition begins. The resultant functional decomposition constrains the categorical options severely; sometimes so much so that the choice for a particular structure is practically inevitable, even if it remains theoretically open. The point is that even though a functional decomposition never strictly entails one unique complete categorical description, anyone with the relevant experience and background knowledge will recognize some functional decompositions to have particular structural descriptions 'written all over them.' I will get back to this point once more below when I recapitulate the role of background knowledge.

How do the three activities just mentioned happen? (1) Specification goes on when functional decomposition is taken in the sense of identifying an overall behavior's mereological parts or when the members of an overall set of behaviors are spelled out. It also happens when functional decomposition has the meaning of identifying a logical or mereological supervenience base. These activities make explicit what was taken for granted in the initial descriptions of the desired behavior. (2) Identifying an overall behavior's parts or its members obviously happens when functional decomposition takes its most literal

meaning of mereological decomposition of behaviors, or, depending on one's preferences in event individuation and composition, when it means identifying the members of an overall set of behaviors. In addition, it can happen through the identification of mereological supervenience bases. (3) Making descriptions of behavior more concrete happens through the identification of causal, conventional, simple, and mereological supervenience bases. Because supervenience bases are always in some sense lower-level, more specific, or more basic than the behavior(s) supervening on them, the transition from supervening behavior(s) to supervenience base always leads to more concrete descriptions of behavior.

The challenging task that follows completion of a functional decomposition is to get from this elaborate ordered collection of concretely described behaviors to a categorical description of an artifact and its components. For some sub-behaviors at the lowest level of the functional decomposition, the choice will be dead obvious since particular components or materials are always used for these behaviors and there is no reason to not follow suit. In other cases, however, a lot of creativity and inventiveness can be required. The functional decomposition might contain behaviors for which no satisfactory solution is known or it might be part of the design assignment to come up with a new solution. In such cases, finding an appropriate categorical description of a component or material will be hard work and there are no guarantees of success. When we assume all of this has been taken care of, i.e., that the designer has managed to come up with possible categorical descriptions of components and materials for every basic sub-behavior in the functional decomposition, the next task might be even more difficult. The designer still has to synthesize all of the individual descriptions of possible structures into physically feasible components and sub-systems of an artifact. Components of artifacts often take care of more than one sub-behavior; the designer has to decide which sub-behaviors to combine into one component. Even if she has categorical possibilities for each of the individual sub-behaviors, that does not mean she has a categorical possibility for a component that combines those sub-behaviors since simply adding up the individual possibilities might be plain inconsistent, extremely hard to realize physically, very inefficient, productive of unwanted externalities, or whatnot. What is more, the same goes at each further level of components: If a designer manages to come up with categorical descriptions of physically feasible components that take care of one or few lowest-level sub-behaviors, these components have to be compatible with the other components with which they will be interacting. They should not adversely interfere with one another. In addition, the components in combination must be affordable, sufficiently efficient, reliable, cheap, fast, safe, silent, noisy, small, big, light, heavy, and whatever else was specified in the design assignment. So it is clearly not the case

that the design process is almost over once functional decomposition is done. Plenty of daunting problems still remain to be solved by a designer.

However, I am now making these problems sound harder than they really are by suggesting that all choices regarding the structure of the prospective artifact have to be made after the process of functional decomposition has finished. I called the assumption behind this idea, viz. that functional decomposition can be done independently of any ideas about structure, (IS) in 4.2.1. In section 4.4.2 I argued against it at length.

Categorical information permeates functional decomposition in at least three ways. First, in the spatial (and perhaps temporal) ordering of the sub-behaviors. Most functional decompositions have an implicit backdrop of assumptions about where the described sub-behaviors are to take place relative to one another. This implies spatial information about the components that will implement those sub-behaviors. In addition, sub-behaviors in functional decompositions are often ordered in time. So if temporal information counts as structural information that is another source of structural information in functional decompositions. Second, the detailed specification of sub-behaviors and their inputs and outputs often strongly suggests categorical information, even if it does not strictly entail it. Choices to specify a behavior or sub-behavior in one or another way will frequently be inspired by ideas about the sort of materials, components or combination thereof are suitable to take care of that behavior in the given design assignment. It is, in fact, hard to see on what other grounds besides such categorical considerations designers could base their decisions to specify sub-behaviors in a functional decomposition. This is one way in which some of the difficult issues about the possible categorical descriptions of an artifact I charted above are (partially) resolved during a process of functional decomposition. Third, particular combinations and hierarchical orderings of behaviors and sub-behaviors are typical for how certain devices and components carry out their behavior. When a designer knows about these devices and components and needs their behavior in a functional decomposition, she may as well decide to use them and decompose the behavior in the way typical of them. In these cases the device or component is already known to be physically feasible, efficient, cheap, etc. Of course, this does not guarantee that the device or component combines well with other components that will be used in this specific design assignment, but at least the thing itself is known to be able to carry out the desired behavior. Again, some of the difficulties outlined above are already resolved during functional decomposition when background knowledge about categorical possibilities guides choices to decompose behavior in one way rather than another.

Hence, with such choices being made during processes of functional decomposition, the amount of categorical information that still has to be

determined when functional decomposition is done is reduced considerably. Sometimes perhaps even to nil, I suspect, but certainly not always. The background knowledge used in functional decomposition may settle some big questions about the operating principles of the artifact and a few of its components, but categorical details will still have to be worked out. And when a design has to be truly innovative and thus cannot rely on standard components and solutions, a lot more than just details requires further attention. Designers will then have to avail themselves of additional methods besides functional decomposition.

With that observation we stumble upon the limits of this chapter and, as it happens, the limits of what I have to say about design processes. My findings unambiguously support the idea that functional decomposition makes an important contribution to the transition from the initial somewhat abstract description of desired artifact behavior to a blueprint, a complete categorical specification of its physical structure. Even though my analysis has at times been critical of the assumptions design methodologists seem to make, its final result should be welcome news as it backs up their advice and provides theoretical foundation for it. To learn more about how designers take the final steps to a complete description of a physical structure, we would have to take a closer look at other methods and techniques recommended by design methodology. Space for that is lacking here; the next chapters are about the second context in which the intentional and physical perspectives on artifacts are brought in contact: explanations of how artifacts can fulfill their technical functions.[43]

[43] Thanks to Tim Lewens for incisive comments on a previous version of this chapter.

CHAPTER V

# Theories of Scientific Explanation

## 5.1 Introduction

This chapter and the two that follow shift the focus away from design towards explanation. Explanations of how an artifact can fulfill its technical function are the second context in which the intentional and physical perspective on artifacts are brought in contact. I briefly explained in chapter 2 that the connection splits in two parts. One part is formed by explanations of how an artifact has a particular disposition, or how it can show a particular behavior. These explanations are purely physical; they do not need intentionality. The other part is a function theory, which explicates which intended behavior is the artifact's technical function and forges a link between this intended behavior and the artifact's structure through the beliefs of a designer or users. When a function theory is superimposed on the abovementioned explanations, we get explanations of an artifact's technical function, i.e., explanations of how an artifact can show the behavior that is its function if it functions well, or of how an artifact is or was supposed to show the intended behavior that is its function if it malfunctions.

We will get to all that in the next two chapters. The present chapter, however, is preliminary in nature. Unlike engineering design, the topic of explanation has been fiercely debated by philosophers over the past six or so decades. To say something about explanation, one must engage with the results of those debates. This chapter does that. It provides an overview of the main accounts of explanation that have been proposed. Because I aim for a reasonable degree of completeness in this, part of the overview is not directly relevant for the specific project of this thesis. I hope the reader will forgive me and appreciate the attempt at comprehensiveness. Those who are familiar with the philosophy of explanation may want to skip directly to the next chapter.

In the course of the overview I will highlight the particular conception of explanation that I will use in the next chapter to construe my account of artifact explanation. I do not intend to suggest that this conception – the causal-mechanical account of explanation – is generally superior over all the other proposals. Such a claim would be highly presumptuous. For my purposes it suffices if the causal-mechanical account is considered by many to be an adequate account of a substantial class of explanations and if it fits the sort of

explanations that I want to model well. So much is beyond doubt, as we will see in due course.

Since the body of literature on explanation is so voluminous I have to limit the discussion in a couple of ways. I make that explicit in the next section. In section 5.3 I present the classification scheme that structures the actual overview of theories of explanation in the remainder of the chapter. The chapter closes with some remarks about the current state of the debate in section 5.8.

## 5.2 Preliminary Remarks and Limitations

Explaining can be a highly diversified activity. A high school teacher might explain a chemical reaction to her class by manipulating little colored balls (representing atoms) and connecting rods (for atomic bindings). A consultant can explain the low morals of a company's staff to its manager by enacting the manager's rude and discouraging behavior towards his employees in a role-playing game. Alternatively, one finds people asserting that science has explained Boyle's law. Or that Einstein's general theory of relativity explains the precession of the perihelion of Mercury. These two sorts of examples reveal an ambiguity in the term 'explanation'. In the former examples, explanation is an act of communication or conveying understanding that might take on any number of forms. The success of such explanatory acts depends on all manner of subjective factors and there do not seem to be objective constraints on the sorts of things that might count as explanations in this sense. No matter what I do or say, if it somehow conveys the right understanding in my audience, it can count as an explanation. The latter examples hint at explanation as an abstract achievement: an ability on the part of scientists to produce some sort of ideal explanatory account of a phenomenon or law. Many people think this ideal is objective; there is such a thing as *the* objectively correct explanation of a phenomenon and this explanation is correct regardless of whether anyone grasps it and recognizes it to be correct. One of the things science does is to produce such explanations. Following Lewis (1986: 217–218), I call the first sense of explanation subjective and the second objective.

Both senses are part of our ordinary use of the notion of explanation, but one might quarrel over the legitimacy or primacy of either. In effect, philosophers favoring an objective, observer-independent account have focused on laying down criteria for what constitutes an ideal explanation, while those favoring a subjective, observer-dependent account have tried to spell out conditions under which a particular act of communication counts as an explanation.[44] Examples of the objective approach are Hempel and Oppenheim's (1948) classic deductive-

[44] Some subjectivists have argued against the very possibility of an objective sense (Van Fraassen (1980) is an example). According to them, explanation is inherently and thoroughly dependent on

nomological (D-N) model of explanation and Salmon's (1984) causal account. Authors such as Van Fraassen (1980) and Achinstein (1983) can be interpreted as working along subjectivist's lines. Personally, I favor an objective account of explanation. In my opinion subjective theories should be understood as really being about the *pragmatics* of explanation. Surely this is an interesting and challenging topic in its own right, but it should not be conflated with the search for an objective account of explanation. In my overview I shall focus on objective theories of explanation. This focus does not exclude the pragmatics of explanation altogether, however, since even the staunchest objectivists will admit that things like the exact content of an explanatory request and the specific parts of the correct explanation that must be given in response to a particular explanatory request are typically determined by context-determined pragmatic factors. Such pragmatic factors do not affect the core of their objectivist claim, however, which is that there are objectively correct explanations of phenomena.

Second, the philosophy of explanation comprises a number of more specialized discussions, which focus on, e.g., functional explanation, intentional explanation, explanation in various special sciences, levels of explanation, and the relation between explanation and understanding. For lack of space it will be impossible to include all of these in my overview. Besides, a lot of these discussions have no clear bearing on the particular type of explanations I am interested in here, so including them would not further the project of this thesis. This chapter thus only presents theories that have been put forward as general accounts of scientific explanation. To the extent that elements of these more specific discussions are relevant, I will draw on them in later chapters.

I should mention the discussion about functional explanation in particular, because that might be thought to bear on the topic of this thesis. The notion of functional explanation has traditionally been used for explanations that seek to explain an item's existence, or persistence, in an encompassing system on the basis of the function of that item (cf. Hempel 1965b; Nagel 1961; Wright 1976; Pettit 1996). In biology, the existence of particular traits of organisms is so explained. In social science, the existence or persistence of social entities or their particular features is accounted for in a similar way. The structure of such explanation is 'item $x$ exists/persists because it has the function to $\phi$.' Functions figure in the explanans; existence or persistence of items is the explanandum. There is an obvious sense in which artifacts are susceptible to functional explanation: They exist because of their functions. Vacuum cleaners

both context and observer. Without arguing the point, I shall assume that both senses are legitimate perspectives on explanation providing complementary insights.

are there because they are for vacuum cleaning. The power button on my laptop computer is there because its function is to allow me to turn my computer on and off. Artifacts are paradigm cases for functional explanation – so much so, in fact, that functional explanations of artifacts are philosophically thoroughly unproblematic.

The major bone of contention in the debate is the scientific legitimacy of functional explanation. Function is a teleological notion and, according to most, the biological realm is not supposed to contain teleology. Nor do people seem to intentionally determine the functions of social institutions, organizations, and the like; many social functions, too, should be construed non-teleologically. Much of the discussion has therefore centered on the possibilities for developing non-teleological notions of biological and social functions. This problem vanishes in the case of artifacts. Their functions are obviously and uncontroversially teleological; they are the intended behaviors for which artifacts are created and used. The same goes for the functions of an artifact's components. They, too, have been created, selected, and put in place because of the behaviors they are supposed to show. As we saw in the previous chapter, a functional decomposition specifies all this. As a result, functional explanation of artifacts is perfectly legitimate and unlikely to stir philosophical controversy.

But even if there had been lots of interesting things to say about functional explanations of artifacts, they would still be irrelevant for the present project. The reason is that I am not looking for explanations that explain the existence of artifacts by means of its function, but for explanations that explain how an artifact can fulfill its function on the basis of its physical structure. My explanandum is not the existence of an artifact or its parts, but how an artifact can show its intended behavior, i.e., can fulfill its function. This is a specific type of explanation and it is not a functional explanation.

This allows me to also make a generalized version of this point. There are lots of things about artifacts that can be explained and doing so requires different types of explanations. As I said, their existence may well be explained by a functional explanation. Alternatively, *intentional explanations* that detail the intentions, desires, and beliefs of their makers may account for artifacts' existence and specific features. People have given *evolutionary explanations* to explain the development of (collections of) artifacts (cf. Basalla 1988; Mokyr 2002; Ziman 2000). To fully understand the workings of technologically advanced artifacts such as lasers or X-ray tubes, one needs *physical* or *chemical explanations*. I could go on but the point should be clear. Artifacts figure in many types of explanations, but I am interested in only one of those here. Of course that does not mean that I think the others are less worthy of philosophical scrutiny. The reason for my focus lies in the broader context of my project.

Explanations of how an artifact can fulfill its function give us a view on the connection between function and structure and that was a central problem in the Dual Nature program.

## 5.3 Classifying Theories of Explanation

Accounts of scientific explanations try to answer one or more of the following questions.

(1) What makes it the case that one thing (an explanans) explains another thing (an explanandum)? Or: What is the nature of the explanatory relation, i.e., the relation between explanandum and explanans?
(2) What are the *relata* of an explanatory relation? Are they, e.g., sentences, propositions, events, facts, or states of affairs?
(3) What are the formal requirements for a good explanation (if any)? E.g., does the explanans have to contain reference to at least one law of nature?

The first question is most directly relevant to the issue of what an explanation is. Hence, I will classify theories of explanation with regard to their answer to this question. The rest of the chapter will present various possible answers, so I shall not give any clarifying examples at this point. Before giving my own classification, however, I confront the one put forward by Salmon (1984). By doing so, I also prepare the way for my own proposal.

### *5.3.1 Salmon's Classification*

Salmon (1984: 15–20, 84–134) distinguishes three conceptions of explanation: epistemic, modal, and ontic.[45] (1) On the epistemic conception, an event is explained by deducing the statement reporting its occurrence from statements about the relevant laws and initial conditions. The explanatory relation holds between statements and is characterized by nomic expectability: the explanandum-statement follows from the explanans-statements with logical necessity. (2) The modal conception takes the explanatory relation to be a relation of physical necessitation[46] that holds between events. The explanans-events could be particular initial conditions being in place and the obtaining of lawful relations between these initial conditions and the explanandum-event. (3) The ontic conception conceives of explanation as showing that an

[45] Salmon introduces his conceptions for the case of a deterministic universe.

[46] Salmon ambiguously speaks of both nomological and physical necessity, the first possibly allowing for other kinds of necessity besides physical. In view of the other conceptions and his talk of 'laws of nature' and 'lawful relations,' I think the modal conception is best construed as having to do with physical necessity.

***Table 2*** *Salmon's Three Conceptions of Explanation*

| | **Epistemic** | **Modal** | **Ontic** |
|---|---|---|---|
| Character of explanatory relation | Logical | Physical | Physical |
| Explanatoriness due to | Logical necessity (deducibility) | (Exhibition of) physical necessitation | (Exhibition of) fit in actual (causal) patterns |
| Character of explanandum and explanans | Statements or propositions | Facts or events | Facts or events |

explanandum-event fits into the discernible patterns of natural regularities in the world. The explanatory relation – which, Salmon claims, is a relation of causality – is the actual relation in the world that holds between the explanans-event(s) and the explanandum-event. This relation is explanatory because it exemplifies a general pattern of 'how the world works.'

The key difference between the modal and the ontic conception seems to be the necessity-requirement. On the modal conception, to explain an event is to show that, given the explanatory events, things could not have been otherwise. For the ontic conception this requirement is too strict; showing how the explanandum-event occupies its place in the natural regularities of the world suffices.[47] Table 2 summarizes the three conceptions.

Let me make some brief comments on why I think Salmon's classification is, on closer inspection, not quite successful. First, both the modal and the ontic conception are ambiguous. Physical necessity may stand for 'necessary in all possible worlds,' in which case an explanation should cite things like essences of objects or processes that necessitate the explanandum. Alternatively, it may stand for 'necessary in all worlds with the same laws of nature as the actual world,' in which case explanation consists in showing a relation of necessity to hold between, on the one hand, certain initial conditions and some subset of the laws of nature and the explanandum on the other hand (Strevens 2004b: 21–22). Another ambiguity cuts through the ontic conception. It is unclear whether Salmon thinks that explanatory force derives from causation alone or from

[47] Note that, in a deterministic picture, this must boil down to demonstrating physical necessitation to hold between explanans and explanandum, as on the modal conception.

pattern subsumption in general. The world contains different sorts of patterns besides causal ones, e.g., mereological, identity, coexistence, or, in more florid ontologies, perhaps even teleological and analogical patterns. Depending on what sort of patterns exist in actual fact, different views of explanation ensue (cf. also Ruben 1993).

Secondly, the characterization of the modal conception hinges on unspecified notions of physical necessity and law of nature. This too may lead to rather dissimilar views of explanation. For example, on a construal in which laws are generalized descriptions of causal relations between objects' capacities, the view ends up very close to a purely causal view – making it more ontic than modal. Alternatively, if laws of nature are construed along positivist lines, the view will become more epistemic. Without a more specific account of laws, the modal conception thus may turn out to be a disguised ontic or epistemic view.

For these reasons, I find Salmon's classification system not very useful, at least not as it stands.

### *5.3.2 An Improved Classification*

The classification principle for my system is how an account of explanation characterizes the explanatory relation.[48] What sort of relation holds between the explanandum and explanans? In virtue of what do some things explain others? Different accounts answer these questions differently and sometimes add specific elements or constraints. The following construals of the explanatory relation capture most of what has been proposed.

(1) *Expectability*: the explanandum is rendered expectable on the basis of the explanans. Given the explanatory facts, the thing to be explained could have been expected to occur.
(2) *Pattern subsumption*: the relation between explanandum and explanans is explanatory because it conforms to a particular well-defined pattern or a member of a class of well-defined patterns. This allows for two different readings:

[48] The source of this idea is Strevens (2004b), but his eventual classes differ from mine. In particular, he does not differentiate between what I call ontic and epistemic pattern subsumption views. He classifies both the particularist and relevance construal as relativistic views. As far as actual proponents of these views exist he is right, but I want to leave open the possibility that they are construed along objectivist lines because I think nothing in the views themselves prevents them from being so developed. Finally, he adds a metaphysical necessity construal of the explanatory relation, in which the *explanans* metaphysically necessitates the explanandum. He attributes such a view to Descartes. I omit that construal here, since it is obviously untenable as an account of the explanations modern science produces.

(2a) *Epistemic*: the explanation derives explanatory force from its exemplification of an epistemic or formal pattern, such as an argumentative structure.
(2b) *Ontic*: explanatory force is due to an explanation's exhibiting some real relation in the world between explanans and explanandum.
(3) *Causation*: the explanans causes the explanandum. Or, more precise, the explanans cites the causal influences, or an appropriate subset of them, on the thing to be explained.[49]
(4) *Relevance*: the explanatory facts are, in a sense to be specified further, relevant to the occurrence of the explanandum.
(5) *Particularism*: there is no saying in advance what the appropriate relation between explanans and explanandum is; it all depends on contextual, pragmatic, social, or other factors. In order not become full-blown subjectivists, prospective proponents of this view should add that, *given* these factors in a concrete situation, it is objectively determinable what the proper explanatory relation ought to be. As far as I know, no one has explicitly defended such a view, although Van Fraassen's (1980) and De Regt & Dieks's (2005) proposals might be reinterpreted as doing so.[50]

It is important not to misunderstand this classification as an attempt to give an exhaustive list of all conceivable construals of the explanatory relation. It is intended to provide an inventory of what philosophers of explanation have in fact proposed. Its only additional claim is that the classes are at least *prima facie* mutually exclusive – with the exception of (3), which is a subclass of (2b) (cf. note 49 above), and (5), which in a way combines (1) through (4). The classes look mutually exclusive at first sight and I see no obvious reasons why more precise renderings of them will inevitably lead to overlap. Of course this does not exclude the possibility of precise definitions that do lead to overlap. For instance, if it is stipulated that causality is one of the ingredients in relevance, then (3) becomes a subclass of (4). Or if you define expectability as some sort of formal epistemic pattern, (1) becomes an instance of (2a). The point is that it is not at all obvious that all possible precisifications of these classes result in overlap, nor even that all plausible ones do.

The following sections add flesh to the bones of this system by discussing concrete examples of accounts that have been proposed in each category.

[49] This view is a subclass of the ontic pattern subsumption view. I put it in a separate class because of its popularity in the literature.

[50] Since this view has no explicit defenders and indeed comes very close to being a subjectivist view, I will not discuss it any further.

### 5.4 Expectability

On the expectability view, an explanation consists in showing how the explanandum was to be expected on the basis of the known facts and relevant laws. Hempel gave the classic statement of this view:

> [A D-N explanation] shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (Hempel 1965a: 337)

Slightly more accurately, Hempel's view can be called a *nomic* expectability view, since the expectability of the explanandum must derive from laws of nature. The characterization holds for both D-N and inductive-statistical (I-S) explanations, only in the latter case we should add that the occurrence was to be expected *with high probability* (Hempel 1962, 1965a). The proper way to express the explanatory expectability relation is obviously an argument to the effect that the phenomenon described by the conclusion had to happen. From this general idea, it follows directly that explanations and predictions have the same basic structure. If we come up with an argument for the claim that a phenomenon had to occur, we could have used that same argument at an earlier time as a prediction that that phenomenon was to occur. Although Hempel claims the D-N model to be an expectability account, it is in fact more of an epistemic pattern subsumption view. That is why I will return to it in the next section.

The I-S model is a clear-cut example of an expectability account. When we lack complete knowledge of deterministic laws governing the behavior of a phenomenon or when such laws do not exist, we can still try to explain the phenomenon statistically. An I-S explanation serves this purpose. It is a sound inductive argument with at least one non-redundant probabilistic law among its premises. Its general form is:

$$\frac{\begin{array}{l} Fi \\ p(F,O) \text{ is high}^{51} \end{array}}{Oi} \text{ [with high logical likelihood]}$$

where $Fi$ stands for a set of factors ($F$) present in this particular instance $i$, $Oi$ for an outcome of type $O$ being realized in instance $i$ and $p(F, O)$ for a probabilistic law to the effect that the statistical probability of co-occurrence of $F$-type facts

[51] How high? At least greater than 0.5 and preferably very close to 1.

and an *O*-type outcome is high. The likelihood of the conclusion, however, has to do with Carnap's notion of inductive or logical probability; it is a measure for the strength of the inductive support the explanans-statements lend to the explanandum-statement. The double horizontal line symbolizes the non-deductive character of the argument.

An example will clarify an additional constraint on I-S explanations. Suppose it is a probabilistic law that 90% of the Dutch people drive a car on a regular basis. Also suppose that fervent environmentalists hardly ever drive cars; only 5% of them sometimes use a car to drive to protest rallies. Tom is a fiery Dutch environmentalist. From the first supposition we can produce a sound inductive argument for the conclusion that, very likely, Tom will drive a car regularly, but from the second supposition we can conclude, with high likelihood, that Tom does not drive a car. Combining the two suppositions eliminates the possibility of constructing any sound inductive argument for either conclusion. For this reason Hempel added the Requirement of Maximal Specificity (RMS) to the I-S model: an I-S explanation should mention all information known to be of potential explanatory relevance to the explanandum. The RMS explicitly relativizes I-S explanations to particular knowledge situations; an explainer's *knowledge of* the facts, as opposed to the actual facts, determines the quality of a given explanation. I-S explanations, then, are after all not purely objective; they are relativized to the epistemic situation of a person or (scientific) community at a certain point in time. Given the general idea of an expectability account of explanation, however, this should not come as a surprise, for the very notion of expectability only makes sense against the background of someone's epistemic situation. What one expects depends on what one knows or believes to be the case.

Let me mention what I think is the most forceful criticism of the expectability view: It confuses reasons for believing that a phenomenon occurred or an argument for the conclusion that a phenomenon occurred with an explanation of that phenomenon (cf. Jeffrey 1971; Salmon 1989, especially §3.6). Expectability is an inherently epistemic notion and whether or not a phenomenon is to be expected by someone is determined by how compelling this person finds the given reasons and arguments. Unfortunately, this need not have anything to do with an explanation of the phenomenon. For that, it seems we need information about the actual facts and events leading up to the occurrence of the phenomenon. If the weatherman reports an oncoming blizzard, then this blizzard is to be expected. But the weatherman's report is nothing like an explanation of the blizzard. The expectability view fails to take into account the explanatory indispensability of (a description of) the actual events leading up to the occurrence of the explanandum.

## 5.5 Pattern Subsumption

Let us now turn to pattern subsumption views. On an epistemic pattern subsumption view, explanatoriness derives from the instantiation of a general epistemic or formal pattern. On an ontic pattern subsumption view it comes from the explanation's describing a real pattern in the world into which the explanandum fits.

### *5.5.1 Epistemic Pattern Subsumption*

Hempel and Oppenheim's D-N model is the simplest example of an epistemic pattern subsumption view. According to the D-N model, an explanation is a sound deductive argument with the conclusion that the phenomenon to be explained occurred. Four adequacy conditions must be met (Hempel and Oppenheim 1948):

(1) Deductive validity of the argument;
(2) Essential reference to at least one law in the explanans;
(3) Empirical content of the explanans;[52]
(4) Truth of the explanans statements (or propositions).[53]

A simple example of a D-N explanation is this:

*Fi*
*F* is always followed by *O*
(or: *F* and *O* always co-occur)[54]
——————————————— [with deductive certainty]
*Oi*

where the second premise is a (deterministic) law of nature. In accordance with the pattern subsumption view, this argument is explanatory because it instantiates the general pattern of deductive arguments satisfying the four adequacy conditions. Or, as Hempel himself phrased it:

[52] This condition is implicitly contained in (1): assuming that the explanandum describes an empirical phenomenon it follows that the *explanans* entails at least one empirical consequence, which is tantamount to having empirical content.

[53] Note that it is not required that the *explanans* contains at least one statement that is not a law. This is because the model needs to account for theoretical deductions of low-level laws from high-level laws as well.

[54] To see the structural similarity with the I-S model even more clearly, substitute the quasi-probabilistic law $p(F, O) = 1$ for this deterministic law.

> [An event] is explained by subsuming it under general laws, that is, by showing that it occurred in accordance with those laws, in virtue of the realization of certain specified antecedent conditions. (Hempel and Oppenheim 1948: 136)

It is now easily seen that the D-N model is really not an expectability view. The fourth adequacy condition states that the explanans statements must be true, not known to be true. Only if someone actually *knows* the premises of the deduction to be true, it follows that she can expect the explanandum-phenomenon. But since the truth of the explanans need not be known, it can hardly be sustained that the D-N model is a genuine expectability account. A given D-N explanation that satisfies the adequacy criteria is perfectly satisfactory regardless of anyone knowing it to be so or anyone ever being in a position to expect the explanandum. An instance of a pattern need not be known or recognized as such in order to be one.

As the first systematic and full-blown account of scientific explanation, the D-N model has been widely debated and numerous criticisms have been leveled against it, countered by ever more sophisticated refinements. One of the most profound problems for the D-N model is its crucial dependence on the concept of a law of nature. Without a yardstick to discriminate between genuine laws of nature and mere accidental generalizations or phony laws it becomes all too easy to conjure up frivolous counterexamples that satisfy the D-N criteria without being explanatory. Suppose it is a contingent truth about our universe that no snub-nosed red-haired philosopher ever wins the lottery. A D-N explanation of Bruce's not winning the lottery could then mention his being a snub-nosed red-haired philosopher. In actual fact, however, Bruce's nose, hair color and profession are of no importance to his not winning the lottery. It seems a correct explanation would have to confer information on Bruce's ticket and on the selection process by which the winning ticket got picked. A more specific version of this problem is what Salmon calls 'the problem of explanatory relevance' (Salmon (1984) ascribes the idea to Kyburg (1965)). We can add all sorts of irrelevant information to a D-N explanation, which makes the explanation highly contestable while it does not violate the D-N conditions. Kyburg's example is that of hexed salt: We can explain that this bit of salt dissolves in water by citing (1) its being hexed and (2) the law that all hexed salt dissolves when immersed in water. The salt's being hexed is utterly irrelevant to its dissolving. Yet the D-N model gives us no grounds to reject this explanation in favor of the correct one that does not mention the hexing.

A second serious shortcoming of the D-N model is its insensitivity to the 'direction of explanation' or the 'asymmetry of explanation' (as first observed by Bromberger 1966). Due to its epistemic and formal character, the D-N model

lacks the resources to discriminate properly between that which explains and that which is explained. If the relevant laws take the standard biconditional form we can simply reverse the deduction, i.e., deduce the initial conditions from the explanandum instead of the other way around. By doing so, we get an argument that satisfies the formal D-N requirements equally well, but is nonetheless a bad explanation. For example, it is perfectly appropriate to explain the red color of a strawberry on a photograph by means of chemical laws, the operational principles of a camera and the fact that the original strawberry was red. But to explain the real strawberry's redness by the same laws and principles, combined with the fact that the picture of it was red, is unacceptable. Explanation is only supposed to go in one direction and cannot simply be reversed. According to many – and this is what motivated the champions of causal accounts of explanation – the lesson to be learned is that explanation ought to track the direction of causation. The D-N model does a terrible job at this.

For a second example of an epistemic pattern subsumption view, let us look at the unification accounts of explanation put forward by Michael Friedman (1974) and Philip Kitcher (1981; 1989). Although they differ in details, these accounts agree in principle on the nature of the explanatory relation. Like the D-N and I-S model, unification accounts maintain that explanations are arguments. According to Friedman, an explanation ought to '[increase] our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given.' (Friedman 1974: 15). For Kitcher (1981), a good explanation is an instantiation of an argumentative pattern that belongs to the set of argumentative patterns that best *unify* our beliefs. Without going into technical details, let me illustrate this approach by an example of Kitcher's. Newton's achievements in explaining the motions of bodies by means of forces acting on them motivated a highly successful scientific program that was characterized by a particular style of mechanical explanation. Newton and his successors attempted to explain phenomena in astronomy, chemistry, optics, physiology, and so forth by casting them in terms of bodies and particles attracting or repelling one another. They did so to discover force laws that govern the behavior of these bodies and particles. According to Kitcher, this procedure is exactly what explanation consists in: finding a few argument patterns that can be used over and over in the derivation of many, diverse phenomena (ibid.: 514). In short, instantiating a unifying argument pattern is what lends an explanation its explanatory force. Explanation becomes a global rather than a local affair; good explanations have a broad scope and can be used for many different phenomena. In consequence, a chief virtue of the unification

account is its aptness for dealing with cases of theoretical explanation (i.e., explanation of laws or principles by fundamental theories).

Elegant though it is, the unification account also has its problems. First and foremost, unificationists have to come up with a sufficiently precise and selective specification of what it means to unify a set of beliefs. Not only must this specification provide a measure to compare different degrees of unification in order to select the patterns that unify *most*, it should also bar cases of spurious unification in a non ad hoc fashion. Suppose that I have a crystal ball in my possession, which, when handled properly, always produces true statements about the world. Now take the following argument pattern.

My crystal ball says it is the case that *p*.
Anything my crystal ball says is true.

---

Hence, it is the case that *p*.

At least at first sight this pattern is highly unifying; it can be used to derive a host of different statements about all sorts of things. Kitcher deals with counterexamples of this sort by stipulating that an argument pattern must be *stringent*, i.e., difficult to satisfy (Kitcher 1989: 433). Stringency is determined by the constraints placed on (1) permissible substitutions for dummy letters in the pattern and (2) the logical structure of pattern instantiations. Since the crystal ball pattern does not place any (non-syntactical) constraints on what *p* can be it is too permissive. Genuinely unifying patterns 'should not be able to accommodate all conclusions' (Kitcher 1981: 529).[55] Assuming that this sketchy criterion of stringency can be made sufficiently rigorous, it remains to be seen whether it is powerful enough to exclude all cases of spurious unification. I will not try to pursue this point here for there is a second challenge that unificationists must face.

Based on a historical survey of several successful attempts at unification of scientific theories, Margaret Morrison (2000) argues that (1) unification does not necessarily lead to explanation nor is it necessarily conducive to explanatory power and (2) a single account of unification cannot be had, since unification has many different faces. The evidence for (2) consists of her case studies and the differences she finds in them. The argument for (1) is slightly problematic. Morrison says that unification does not always give us insight in the causal mechanisms underlying the phenomenon to be explained. In fact, sometimes it even does the exact opposite by clouding our insight in the underlying

[55] This demand is similar in vein to the D-N requirement that the *explanans* have empirical content. Both aim to exclude 'explanations' that can explain everything.

mechanisms. This implies that the unification account of explanation fails Morrison's test, since she stipulated that the aim of explanation is to exhibit the causal mechanisms responsible for the production of explanandum (Morrison 2000: 3, 31). In my opinion this argument is unpersuasive. Claiming that the unification account of explanation is unsuccessful because it is not a causal account begs the question, since its whole point is to provide an alternative for causal accounts of explanation.

The argument can be repaired by remembering the point about the asymmetry of explanation. Why is the sort of pattern that explains the color of a strawberry with a deductive argument involving the color of that strawberry's photograph less unifying than the pattern that does it the other way around? Kitcher counters that such a set of patterns 'is less good at unifying our beliefs than the set we normally choose for explanatory purposes' (Kitcher 1981: 525). But how obvious is this? We can think of lots of similar patterns that all get the direction of explanation wrong, but together would constitute a pretty unifying set of patterns. I am therefore not very confident that the unificationist can be let off the hook that easily.

Let me elaborate on this point. There is a deeper trouble underlying the asymmetry problem and this trouble pervades all epistemic pattern subsumption views. The intuition that makes the asymmetry problem so appealing is that explanations should somehow mirror real states of affairs and processes in the world. They should not just be formed and constrained by epistemic and formal criteria. To put it crudely, there is a strong intuition that the ultimate criterion to distinguish good explanations from bad ones should be provided by the world out there and not patterns in our knowledge of that world. As long as epistemic pattern subsumption views remain faithful to their epistemic character, meeting this markedly ontic demand will be outright impossible.

Before we go on to examine ontic pattern subsumption views, which fare better with regard to this problem, I want to present one more epistemic pattern subsumption view: Cartwright's *simulacrum* account of explanation. According to this view, 'to explain a phenomenon is to find a model that fits it into the basic framework of the theory and that thus allows us to derive analogues for the messy and complicated phenomenological laws which are true of it' (Cartwright 1983: 152).

The explanandum is explained by 'seeing' it through the conceptual, mathematical, and logical framework of a scientific theory and this seeing is accomplished by creating a model that subsumes the explanandum under the general pattern laid out by the theory.[56] That this pattern is indeed epistemic in character follows from Cartwright's firm assertion that '[a] model is a work of

fiction' (ibid.: 153). Models need not be realistic in the sense that they picture the phenomena accurately. What makes a model successful *qua* explanation is its ability to replicate much of what goes on precisely, but this is highly dependent on pragmatics, since 'models serve a variety of purposes, and individual models are to be judged according to how well they serve the purpose at hand' (ibid.: 152). These purposes may include a precise calculation of some quantity, an exact rendering of a functional relationship, or coarse replication of a broad range of behaviors, or insight in the causal chain leading up to some event.

I have two short comments on this view. First, it seems to me that it is also susceptible to the asymmetry problem, since there are no clear ontic constraints on the contents and structure of an explanatory model. At least at first sight, models that get the direction of explanation backwards may still be able to replicate a lot of what goes on yet we would hesitate to call them truly explanatory. Second, the simulacrum account is intimately bound up with Cartwright's overall picture of science and indeed hardly makes sense if one does not also accept her anti-realist construal of laws, theories, and models. Sure enough, this is not in itself an argument against the simulacrum account, but in general it does not seem very desirable for a theory of explanation to require that one automatically buys into a package of other interconnected anti-realist views about science. One needs good reasons beyond those related to explanation to do so.

### *5.5.2 Ontic Pattern Subsumption*

The general idea of ontic pattern subsumption views of explanation derives straight from Salmon's ontic conception discussed above: 'To explain an event is to exhibit it as occupying its place in the discernible patterns of the world' (Salmon 1984: 18). Interestingly, however, in some moods Hempel seems to have had something like it in mind as well: 'The understanding [explanation] conveys lies ... in the insight that the explanandum fits into, or can be subsumed under, a system of uniformities represented by empirical laws or theoretical principles' (Hempel 1965a: 488).[57] Salmon immediately goes on to limit the class of explanatory patterns to causation only, but I think his general idea allows for a more permissive position. On this view, a successful explanation shows how the explanandum fits into real patterns in the world. This exhibition

[56] Cartwright does not use the word 'model' in its strictly logical sense, but in a general, scientific sense, as a term for any formal construct that represents a behavior pattern.

[57] If we add an ontic twist to Kitcher's argument patterns (e.g., that argument patterns have to correspond to real patterns in the world), unificationism too starts to resemble an ontic pattern subsumption view. If my Kitcher exegesis is correct, however, this would fly in the face of the spirit of his account.

may of course take the form of an argument, but it need not. An informal description, an enumeration of causal influences, or even a picture may be just as adequate a means.

An urgent question for this view is: What sorts of patterns populate our world and are they all suited for giving explanations? If we put aside Humean doubts about its reality, the first and most obvious contender is causation. Yet I find it hard to see why we should limit the set of possibly explanatory patterns beforehand to causation only, as Salmon does (the next section discusses his and like-minded approaches in more detail). Of course, it might turn out that in actual fact causation is the sole source of regularity in our world, but *a priori* it seems perfectly possible that the world is full of other than causal patterns. This is partly an empirical question, to be decided by how well science manages to describe and explain phenomena by means of causal and possibly other patterns. Let us nonetheless look at other contestants.

A second candidate is identity. Besides being ontologically even less suspect than causation, identity even seems to function as an explanatory pattern in some accepted scientific explanations. Take the explanation of the temperature *T* of some gas sample by the mean kinetic energy of its constituent molecules *E* (see Achinstein (1983: 233–237) for more examples of this sort). Since (1) *E* and *T* are the same property (for a gas to have a particular temperature simply is for its molecules to have a particular mean kinetic energy), (2) *E* explains *T*, and (3) nothing can cause itself,[58] the explanation cannot be causal, so we seem to have hit upon a good explanation that works by subsuming its explanandum under a non-causal ontic pattern, namely identity.

Are there still other ontic patterns that might function in an explanation? Supervenience may be a third option to consider. I do not know of any uncontroversial examples of supervenience, but various authors have proposed supervenience relations between (1) biological properties (e.g., being alive) and physicochemical properties, (2) mental properties (believing, feeling, or intending something) and physical properties, (3) functional properties (such as being a screwdriver or being a beautiful painting of a landscape) and an object's physical and chemical make-up, and (4) dispositions and their categorical bases. If such proposals are correct, we may have to allow for explanations that subsume their explanandum under a supervenience pattern. An explanation of sugar being dissolvable in water in terms of the chemical composition and properties of sugar molecules might be an example of this. So *if* supervenience is indeed a distinct ontic pattern, ontic pattern subsumption views ought to consider it as yet another possibly explanatory pattern.

Kim (1974) suggests that even this provisionary list does not exhaust the

[58] Excluding God or other supreme beings, this seems a reasonable principle to accept.

supply of patterns in the world. He adds the following non-causal relations: (1) Cambridge dependency (a relation between two different events that have the same material basis, e.g., Socrates's death and Xantippe's becoming a widow), (2) the by-relation for actions and events (one action being done by means of another, e.g., starting a car by means of turning the key), and (3) event composition (a temporal part-whole relation defined for events, e.g. typing a sentence as part of writing a paper). We might further want to add (4) mereological relations for physical objects and (5) relations of coexistence or exclusion. Perhaps even this much is not enough. Although the existence of some of these patterns is contentious and strongly dependent on other contestable ontological claims about for instance the identity of events, it is by no means evident that we can plainly dismiss the entire list.

Where does this put prospective proponents of an ontic pattern subsumption view of explanation? Two options are open. The first is to accept a plurality of ontic patterns and concede that all are sources of possible explanations. As a result of this explanations might come in very different sorts, depending on the ontic pattern under which the explanandum is subsumed. Some explanations will be causal, others mereological and yet others will use supervenience. This is not very parsimonious, but then again, why should we expect the world be particularly attentive to our desire for simple and parsimonious theories of explanation? This pluralist option has to face a more important problem. What if an explanandum-phenomenon occupies a place in two or more different types of patterns? A simple example can be constructed by considering a solution with a changing pH-value. We can explain this change with an 'identity-explanation' by pointing to the fact that the concentration of hydrogen ions in the solution is changing. But as it happens we can also explain this change causally by describing the chemical reaction that is taking place because we have added a sample of sodium carbonate. Does that mean that there are two concurrent types of explanations of the same phenomenon which are both true? If so, is one better than the other or are they both equally satisfactory? I think most people would be intuitively inclined to regard the causal explanation as superior, but there is nothing in the general idea of an ontic pattern subsumption view to corroborate this preference. This consideration raises the question of whether different ontic patterns perhaps have different degrees of explanatoriness. For all I know, nobody has tried to answer this question and a discussion of it is also beyond the scope of this overview.

The second option for the defender of an ontic pattern subsumption view is to assert that only one pattern is suited for explanatory purposes. The most obvious candidate for this position is causation since it is by any standard the most ubiquitous pattern from our list. Since this choice has been highly favored and, consequently, developed to an impressive degree of sophistication, the

next section will be completely devoted to it. One question remains for this second option: What is it about causation that makes it so happily wedded to explanation, in contrast to other patterns? Is explanation somehow intrinsically connected to causality, so that demands for explanation are only answered satisfactorily by giving causes (as opposed to other ontic patterns)? This seems to be the position implicitly taken by many advocates of causal views, but they hardly argue for it.[59]

The reader may wonder if anyone has ever defended a general ontic pattern subsumption view of explanation. Well, not that I know of. It remains an open question whether such an account is feasible (cf. Ruben 1993), although we will see below that the mechanistic variety of causalism incorporates elements of a more general ontic pattern subsumption view.

## 5.6 Causation

The basic position of causal accounts of explanation is as simple as it is powerful. Causality is the stuff that glues our world together and in order to explain a phenomenon we have to show how it came about by describing the causal influences that led to its occurrence. Spelling out this basic idea has proved difficult. I will discuss different causal accounts; along the way, their intricacies and problems should become clear.

### *5.6.1 Salmon's Causal account of Explanation*

The seminal figure for the causal view of explanation is Wesley Salmon. His thoroughly developed view is part of the explanation orthodoxy, just like the D-N model. Let me try and present it concisely. Salmon summarizes his own view as follows:

> Explanation is a two-tiered affair. At the most basic level, it is necessary … to subsume the event-to-be-explained under an appropriate set of statistical relevance relations … At the second level [these relations] must be explained in terms of *causal* relations. The explanation, on this view, is incomplete until the causal components of the second level have been provided. (Salmon 1984: 22)

[59] There is a third conceivable option for ontic pattern subsumption views. One could argue that really there is only one ontic pattern in the world and that other patterns are reducible to this pattern. This seems highly implausible to me; at least identity and causation are separate irreducible ontic patterns. Therefore I think that this third option is not tenable and that my point survives. Advocates of the causal view are bound to defend their claim that only causation leads to proper explanation.

Explanation thus comes in two parts; first the explanandum is placed in a system of statistically relevant factors[60] and then this system is underpinned by a causal story that shows how the statistically relevant factors work together to cause the explanandum. Since Salmon's official presentation is rather formal and abstract I will further introduce his account by means of an example that contains the principal ideas. Suppose we are testing airplane wings of various sorts and want to explain why a particular airplane wing broke. This first phase consists of constructing a statistical relevancy (S-R) basis. The first step in this phase is to determine the appropriate contrast class for the explanandum, e.g., why did it break rather than not break, or why did it break rather than buckle or warp? Next, we determine the set of statistically relevant factors, for instance the shapes of the wings, their producers, the materials that they are made of, how long they have been tested so far, and so forth. The account requires that this set of factors is complete; it should be impossible to identify further relevant factors that affect the probability of the wing's breaking. Then prior probabilities need to be ascertained for the events of breaking and not breaking, and posterior probabilities relating to every factor, i.e., probabilities of breaking and not breaking, given that it has shape $s_i$, is made of material $m_i$ by producer $p_i$ and has been tested for $t_i$, and so forth for all shapes, materials, producers, and testing periods. Should it turn out that one of these factors does not affect the probability of breaking (e.g. different producers' wings all have the same probabilities of breaking under equal circumstances), then this factor should be removed from the list. In doing so, we make sure to employ *all* relevant factors and *only* those that are relevant. Finally, we find the specific class to which our particular wing belongs (shape, producer, etc.) and that gives us the S-R basis for the explanation.

The second phase then provides the causal underpinning of the S-R basis. To this end, Salmon develops an original theory of causality (1984, especially chapters 5 and 6). Instead of the traditional picture of events causing one another, he sketches an ontological picture in which *causal processes* transmit causal influences. A causal process is distinguished from a pseudo-process by the *criterion of mark transmission*: A causal process is capable of transmitting marks, a pseudo-process is not.[61] A single intervention at one point in a causal process transforms some characteristic of the process in a way that persists

[60] Factor *C* is statistically relevant to the explanandum *B* under circumstances *A* if and only if $P(B|A.C) \neq P(B|A)$. Example: Why did this heated glass *(A)* crack *(B)*? Because it was put in ice-cold water *(C)*. Being put in ice-cold water is statistically relevant because the odds of the glass cracking are altered considerably by it.

[61] Note the counterfactual twist in this formulation. A process need not actually transmit a mark; the *capability* to do so suffices.

from that point on (provided there are no additional nullifying interventions). A driving car is a causal process, the movement of its shadow a pseudo-process. The next part of Salmon's account of causality is the notion of *causal interaction*. The following quote gives an informal characterization and also shows its close connection with mark transmission:

> Modifications in processes occur when they intersect with other processes; if the modifications persist beyond the point of intersection, then the intersection constitutes a causal interaction and the interaction has produced marks that are transmitted. (Salmon 1984: 170)

Causal processes propagate structure and order throughout the world; causal interactions produce modifications in that structure and order. There is yet another source of structure and order, the so-called *conjunctive common cause*: a situation in which two independent and not directly interacting causal processes arise out of a special set of background conditions and therefore exhibit an otherwise highly improbable degree of correlation among their effects. For instance, two fraudulent students copying their paper from the same original. Together, these three notions (causal processes, causal interactions, and conjunctive common causes) make up the 'causal structure of the world.' To complete the second phase of explanation, the S-R basis should be accounted for by a description of the causal processes, causal interactions, and conjunctive common causes that bring about the ascertained probabilities.[62] While this final step may not always be wholly possible in practice, it should serve as an ideal.

This, then, is Salmon's ingenious account of causal explanation. Much more can be said about the technical details of its central concepts, but I think it more instructive to look at three more general issues. First, why do we need the S-R basis? Once we have told the causal story and laid bare the mechanisms by which the explanandum was produced, what more could we want and why would anyone need the S-R basis any longer? We stumble upon a dilemma here: Either the S-R basis is an essential ingredient for successful explanation, or it is of mere heuristic value in yielding the causal story. The first horn leads to an idiosyncratic picture of explanation with two more or less separate components. There is nothing wrong with idiosyncrasy as such, but Salmon has given us no

[62] Salmon has little to say on this last step. Perhaps he believes it to be unproblematic. It seems to me, however, that it is far from trivial to explicate an elaborate system of S-R relations in terms of underlying causal processes and interactions. To mention a few worries, how do presumably deterministic causal processes translate into probability relations and how do the S-R relations between events and/or factors tie up with the underlying process ontology?

arguments as to why we should accept this particular picture. The second horn makes one wonder what reason there is for expecting that S-R relations are easier to come by than causal ones. After all, there must be some sort of benefit in using the S-R basis as evidence for the causal part of the explanation if the S-R relations are to have heuristic value. In fact, I think the S-R basis may actually lead one astray in finding the right causal account, especially in the case of back-up causes and preemption. Suppose we are still explaining the airplane wing's breaking. Also suppose that when the wing was about to crack, a heavy piece of metal fell on its edge, causing the wing to break. In this case, the S-R basis would need to be supplemented with probabilities of the wing's breaking, given that this piece of metal hit it, but the basis would also still include the shape, material, producer, and testing times. Yet this is clearly redundant or even wrong; in this particular case, the correct causal explanation of the wing's breaking is simply the fact that a piece of metal of such-and-such proportions and weight fell on it. The other S-R factors do not belong in an explanation of this particular instance. I conclude that statistical relevance is not always a reliable device for picking out the right causal factors; statistical relevance is no guarantee for causal influence. Perhaps this is not so much a criticism as a suggestion to purge the S-R part of the account.

Secondly, does explanation really require delineating *all* statistically relevant factors and *all* causal processes and interactions behind them? Apparently, changes in solar activity (outbursts that result in electromagnetic storms in our solar system) have an effect on the climate of our planet, thus influencing the growth of agricultural crops. Now suppose we want to explain why farmer Bob's corn harvest was significantly lower than last year's. As it happened, Bob developed an obsessive interest in truck races and spent a substantial part of summer not looking after his corn, but practicing his racing skills on his cornfields. Given the statistical relevance of solar activity, the Salmonian S-R basis for explaining Bob's diminished harvest would need to include probabilities related to solar activity and climate change. But again this is clearly wrongheaded, for even if the solar activity had been excessively beneficial to corn growth, Bob's harvest would still have been deplorable. Growing corn and racing trucks on the same fields just is bad farming and beneficial solar activity could not have done anything to prevent the decreased harvest – it is simply irrelevant in this case. So apparently not every causal influence is always explanatorily relevant.[63] We need some sort of criterion of explanatory

[63] It is of course possible to maintain that a complete explanation must mention each and every causal influence on the explanandum, but this seems at odds with ordinary (scientific) explanatory practice. We do not care for all causal influences, we want to know which ones were decisive and really made a difference. To maintain that an 'ideal' explanation should nonetheless consist of a enormous

relevance and both the criteria of being statistically relevant and exerting causal influence are too liberal. Salmon's problem of explanatory relevance (cf. section 5.5.1) bites back (Hitchcock 1995).

Counterexamples form a third concern for Salmon's account: (allegedly) non-causal but nonetheless accepted scientific explanations. Since such cases impinge on all causal views of explanation, I will discuss them at the end of this section. It is now time to look at other causal accounts.

### *5.6.2 Other Causal Accounts*

Another classic exponent of the causal account of explanation is Railton (1981).[64] His view is very close to Salmon's except for the S-R part. Here is a quote:

> An ideal text for the explanation of the outcome of a causal process would look something like this: an interconnected series of law-based accounts of *all the nodes and links in the causal network* (my italics, JdR) culminating in the explanandum, complete with a fully detailed description of the causal mechanisms involved and theoretical derivations of all the covering laws involved. ... The actual ideal is not to *produce* such texts, but to have the ability (in principle) to produce arbitrary parts of them. (Railton 1981: 247)

Given this striking similarity, it is obvious that Railton's conception suffers from the same explanatory relevance problem. Apparently he hopes to solve this problem by an appeal to pragmatics: Contextual factors determine what part of the ideal text is a good explanation. Railton is not very explicit here, however, so it remains unclear how he thinks this is supposed to work. I will get back to this point in due course, since the next account to be discussed works along the same lines and is somewhat more explicit. Additionally, Railton does not analyze the notion of causality itself, which leaves us with the question of what exactly qualifies as a 'node and link in the causal network'. But then again, giving an account of causal explanation and giving an account of causality need not necessarily be the same project.

David Lewis's account of explanation is more sensitive to the problem of explanatory relevance and gives a possible solution to it. His starting point is again very similar in vein to that of Salmon and Railton:

story about each and every tiny causal influence leading up to the explanandum seems to me undue philosophical arrogance.

[64] Perhaps this is not quite right, since Railton holds that not all explanations need to be causal. He distinguishes causal from probabilistic explanation.

> [To] explain an event is to provide some information about its causal history. ... Among the true propositions about the causal history of an event, one is maximal in strength. It is the whole truth on the subject – the biggest chunk of information that is free of error. We might call this the whole explanation ... It is, of course, very unlikely that so much explanatory information ever could be known, or conveyed to anyone in some tremendous act of explaining! (Lewis 1986: 217)

If an act of explaining is not an attempt to give this whole explanation, we may wonder what determines the quality of explanations if it is not the degree to which they approximate wholeness. Like Railton, Lewis is not very clear about this. Explanatory information may concern various parts of the causal history and include a more or less detailed description of causal mechanisms, a description of causal chains of events, an existential statement about the causal history including such-and-such an event, mention of some global property of the causal history, or even negative information about what the causal history is not or does not contain. It all seems to depend on a pragmatic and context-dependent criterion of saliency and the knowledge state of the explainer (some people just know more about an explanandum than others and can therefore provide more detailed explanatory information).

At first blush, this might seem to be an elegant solution to the explanatory relevance problem. Nonetheless, I have a serious objection against it. Since Lewis is more explicit than Railton, we can see that this solution in effect boils down to abandoning the ideal of an objective account of explanation. As long as some information about the causal history is conveyed, anything goes in principle. Explanatory goodness depends solely on pragmatic, context-dependent, and subjective criteria. I think the intuition that some explanations are objectively better than others is too strong to throw in the towel so soon; we should at least try and come up with a more objective criterion of explanatory relevance. To spell out this intuition, imagine that context and pragmatics alone determine which parts of the causal history are relevant. Then it should be possible in principle to think of contexts, however extraordinary, in which seemingly irrelevant parts of causal histories become good explanations. For instance, contexts in which the solar outbursts become part of a good explanation of Bob's meager harvest. Or contexts in which the gravitational effect of Mars is part of an explanation of the breaking of the airplane wing. Many people, myself included, find this hard to believe. Therefore, we will now look at two philosophers who have tried to give an objective criterion for explanatory relevance.

Lipton's analysis of contrastive explanation (Lipton 1991) can be interpreted as a solution to the problem of picking out the explanatorily relevant causes from the complete causal history.[65] He proposes that most or perhaps all instances of causal explanation are *contrastive*. What needs to be explained is not a fact or event on its own, but why this fact obtained *rather than* some other fact, or why this event occurred rather than that. How does that help? Contrastive questions guide the selection of explanatorily relevant causes by means of the *Difference Condition*: 'To explain why *P* rather than *Q*, we must cite a causal difference between *P* and not-*Q*, consisting of a cause of *P* and the absence of a corresponding event in the history of not-*Q*' (ibid.: 43).

To explain why farmer Bob's, rather than his neighbor's, harvest was so miserable, we should look for a causal difference between Bob's farming and that of his neighbor. In this case, that would obviously be Bob's unsavory racing ambitions, which his neighbor did not share. There may be several causal differences at work in one and the same case. Besides speeding through his cornfields, Bob might also have spent quite some time watching videotaped truck races when he should have been working his fields. The difference condition does not entail that there is only one way to explain a contrast. All differences may be explanatory in particular contexts; which ones are right is governed by various principles. For instance, we tend to prefer counterfactual differences if these are available: Cases in which, had the corresponding cause been present, *Q* rather than not-*Q* would have occurred.

This is indeed a promising approach, but it does raise a question: How do we determine the appropriate contrast in a given request for explanation? After all, not every explanatory request is explicitly formulated as a contrastive question. Lipton contends that this should be settled by a pragmatics of explanation. It depends on contextual factors such as the speakers' (fore)knowledge, the history leading up to the request for explanation, the institutional context in which the request takes place, etc. Even so, for Lipton's theory to hold in general, we need to know whether every explanatory question can be plausibly rephrased as a contrastive question. One could argue that we can always rephrase 'Why *P*?' as 'Why *P* rather than not-*P*?' for any *P*. An appropriate causal difference between *P* and not-*P* must then be something like a sufficient cause for *P* (since that is by definition not in the history of not-*P*).

This brings us close to a second solution to the problem of explanatory relevance. It is due to Strevens (2004a). According to him, the fundamental laws of physics govern causal influence. So in order to get the causal history of the

[65] I should add that Lipton is not an exclusive causalist; he thinks lots of good explanations are causal, but not that all explanations must be causal.

explanandum straight, we should, in the end, consult these laws. He thus defers part of the development of an account of causation to the physicists.

How do we then select the relevant explanatory causes from the causal history? The basic answer is remarkably simple: only those causal influences that *make a difference* to the explanandum are explanatory. 'The explanatorily relevant parts of any causal network are the elements that made a difference to whether or not the explanandum occurred' (ibid.: 158). The set of difference makers is selected from the total set of causal influences by means of an eliminative procedure (ibid.: 162–164). Given a correct and complete description of a causal history, we gradually remove those causal influences from the description that do not eradicate the causal production of the explanandum until nothing further can be removed. What remains is a set of difference makers for the explanandum. Jointly, this set will constitute a sufficient cause for the explanandum.[66]

This basic answer is not adequate for a number of reasons, one of them being that the procedure is too crude. Either a causal influence is completely removed, or it is not. The procedure does not allow for partial removals. In our airplane wing example, this would mean that the exact mass and volume of the piece of metal that caused the breaking count as difference makers, whereas in reality only the fact that it was a fairly heavy and voluminous piece of material falling with a roughly defined speed should count. To remedy this, Strevens refines the account with an abstraction operation (ibid.: 167–170), which makes it possible to remove causally irrelevant specifications from difference makers. There are more refinements, one of them being an explication of how this account deals with (theoretical) explanation of laws, which is another traditional problem for causal accounts. The technicalities, however, are beyond the scope of this overview.

The final branch of the causalist family consists of people who have picked up on and developed the notion of a mechanism. The Railton quote above characterizes explanation in general terms as the unearthing of causal mechanisms and Salmon describes his view in similar terms at various points. The idea that explanation involves the uncovering of mechanisms is in fact much older and dates back to the natural philosophies of 17th century scientists like René Descartes, Christiaan Huygens, and Sir Isaac Newton, who conceived of nature as a grand mechanism to be analyzed by science. What all of these people failed to do, however, is to give a more precise account of what they take a mechanism to be; that notion is left as a primitive. This has recently led a number of philosophers to develop more precise conceptions

[66] At this point, the similarity with Lipton's proposal clearly shows.

of what a mechanism is (Bechtel and Richardson 1993; Glennan 1996, 2002; Machamer *et al.* 2000; Craver 2001). We will make closer acquaintance with their work in the next chapter because the account of artifact explanation that I develop there is mechanistic in character. I limit myself to a brief and general outline of mechanistic ideas here. Proponents of a mechanistic view of causal explanation take seriously the idea that explanation should describe the underlying mechanism productive of an explanandum-phenomenon and argue that more is involved in describing a mechanism than the mere citing of one or more causal influences that cause the explanandum. Mechanisms are more or less complex systems of interacting objects. They are found everywhere in the natural and artificial world: cells, organisms, social groups, watches, etc. To describe a mechanism one should focus on the objects that are involved in one of its behaviors. A description of these objects, their parts, their material constitution, and the ways in which they interact dynamically should make clear how the phenomenon that is being explained occurred. Depending on the precise explanatory request the mechanism must be described in more or less concrete and detailed terms.

A few remarks about mechanistic explanation. The point that causal explanation involves the detailed description of a productive mechanism does not exclude the solutions I discussed above to the problem of explanatory relevance. There is no reason why mechanists could not embrace a solution like Lipton's or Strevens's to this problem. I think the mechanistic point is best interpreted as an addition to, and not a rejection of, the general causal accounts I discussed before. Next, by stressing the importance of (a description of) the objects involved in causal production, this mechanistic view in effect builds in elements of a more general ontic pattern subsumption view. By describing objects and their parts and material constitution, one has to take into account mereological relations, identity relations, and perhaps also supervenience relations. Finally, I think the mechanistic view is most helpful for the project of this thesis, i.e., the development of an account of artifact explanations that explain an artifact's ability to perform its technical function on the basis of its physical structure. Therefore, the account I develop in the next two chapters draws on resources from the mechanistic tradition.

In closing this section, I want to bring up the problem of non-causal counterexamples. Various authors have suggested that science provides us with several widely accepted non-causal explanations. I will mention a few of them without going into technical details. There are explanations in mathematics and philosophy that do not cite causes, and there is an ongoing debate as to whether intentional explanation is causal. Kitcher (1985: 636–637) discusses the explanation of why neon is chemically inert by quantum chemistry and

various explanations in formal linguistics. Cartwright (1983: 21) talks about 'laws of association' that figure in physical explanations. For instance the law that whenever the force on a classical particle of mass *m* is *F*, its acceleration will be *F*/*m*. Glymour (1982: 184–186) presents examples in which a variational principle explains gravitation and electro-dynamics. He claims that although these explanations may have causal ramifications, that is not what makes them explanatory. Railton (1978: 207) says that some facts are explained non-causally by subsumption under structural laws such as the Pauli exclusion principle. And finally, Sober (1983) distinguishes an entire class of what he calls 'equilibrium explanations' in biology, which, he maintains, are not causal. Space is lacking to go into counterarguments from the causal camp, but let it suffice to say that all of these (alleged) counterexamples have been contested by adherents of the causal view.

This ends my discussion of causal theories of explanation. The short remainder of the chapter will look at a final construal of the explanatory relation as a relevance relation.

## 5.7 Relevance

Some have conceived of the explanatory relation as a relation of relevance: the explanans consists of facts or events that are relevant to the occurrence of the explanandum. The exact nature of such accounts of course depends on the specification of this relevance relation. I will consider two proposals, the first by Salmon and the second by Van Fraassen.

In earlier work, Salmon (1971) put forward the statistical relevance part of his later causal theory of explanation as a self-contained account of explanation. In this account, to explain an event is to place it in a system of relevant statistical generalizations in very much the same way as was discussed above. Since the problems with this proposal will be evident from the earlier discussion, there is no need to dwell on them here. First, statistical relevance is not always a reliable indicator of explanatory relevance. The fact that, on average, certain events tend to have an effect on the probability of an explanandum occurring does nothing to show that they had that effect in this particular instance. Second, the criterion of statistical relevance is too liberal (i.e., it lets in too many explanatory factors) and therefore suffers from the modified problem of explanatory relevance that also troubled the causal account of Salmon, Railton, and Lewis.

Another proposal is due to Van Fraassen (1980). According to him, an explanation is an answer to a why-question asked in some context. A why-question *Q* has the general form 'Why $P_k$?', where $P_k$ is the explanandum. *Q* can be analyzed as an ordered triple $<P_k, X, R>$ where *X* is the contrast class $\{P_1, \ldots, P_k, \ldots\}$ and *R* a relevance relation. (As in Lipton's account, the question becomes 'Why $P_k$, rather than the other $P_i$'s?') An explanation is an answer to *Q*, which

has the following canonical form: '$P_k$ in contrast to the rest of $X$ because $A$,' where $A$ bears relation $R$ to $<P_k, X>$. Answers are to be evaluated according to (1) the correctness or plausibility of $A$ itself, (2) the extent to which $A$ favors $P_k$ as against the other members of $X$, and (3) how well $A$ does in comparison with other possible answers to $Q$.

The relation between explanandum and explanans (between $Q$ and $A$) is a relevance relation, but that is not saying much. What sort of relations might qualify as relevance relations? What is it about a prospective $A$ that makes it bear a relation of relevance to $<P_k, X>$? Remarkably, Van Fraassen does not put any constraints on $R$ so that, as Kitcher and Salmon (1987) cogently argue, the account is trivialized since virtually any proposition can count as an answer to virtually any explanatory question, which is of course utterly undesirable – not to mention false. Although we do find occasional remarks on the sort of relations that can be relevant to explanatory questions (e.g., physical necessitation, causality, intentional relevance, fulfilling a function, etc.), Van Fraassen does not translate these remarks into requirements on $R$. He does emphasize that the context of the question determines explanatory relevance. So in the end Van Fraassen's stated account turns out to be a full-blown relativist construal of the explanatory relation. The only way to alter this predicament is by placing constraints on what sort of relations can be relevant and hence explanatory, but that is exactly what the theories of explanation I discussed in this chapter have tried to do. In sum, then, Van Fraassen would still need to solve most, if not all, of the traditional problems that have plagued other theories of explanation.

## 5.8 Concluding Remarks

To conclude the chapter, I want to make an attempt at describing the current state of affairs in the philosophy of scientific explanation. Inevitably, this attempt will reflect my personal perspective but I will try to be as objective as I can.

First, I think it is safe to say that Hempel's D-N and I-S models of explanation have acquired the status of relic. Highly valued for sure, but relics nonetheless. The intuitions underlying Hempel's account have been picked up and elaborated by pattern subsumption views (both epistemic and ontic) but, as far as I know, the strict formal constraints and even the expectability construal of the explanatory relation have lost their currency.

Secondly, today's orthodoxy comprises both the unificationist and the causal views of explanation. Both views are considered to be pretty good theories of scientific explanation, although everybody is aware of their respective problems. Various philosophers have recently defended their own versions of unificationism (Schurz and Lambert 1994; Weber 1996; Jones 1997; Schweder

2004) or causalism (Humphreys 1989; Woodward 1997, 2003; Ylikoski 2001; Strevens 2004a, b). Interestingly, both Kitcher and Salmon have, in spite of strong earlier claims to the contrary, suggested that a viable account of scientific explanation might have to include both causal and unificationist elements (Salmon 1989; Kitcher 1989). In fact, some of the authors I just mentioned indeed try to develop something of a synthesis between both views; unificationism with a causal twist or causalism with unifying virtues. There seems to be a general agreement that causal accounts are especially apt for the explanation of singular events, whereas unificationism intuitively does a better job at theoretical explanation. Hence, the main challenge for proponents of either account is to show that their favored version of causalism or unificationism can handle theoretical explanation or explanation of singular events respectively. The other construals of the explanatory relation have not received the broad support of the unification and causal accounts, but are nonetheless part of the received wisdom of the philosophy of explanation.

Thirdly, to some extent the debate over explanation is tied up with the realism vs. anti-realism controversy. Anti-realistically minded philosophers tend to prefer subjective or relativist accounts of explanation, whereas realists generally opt for a more objective account of explanation. Hence, Van Fraassen's and Cartwright's views of explanation do have currency in anti-realist circles. Realistically inclined philosophers of science, however, favor a causal or unification view.

Finally, it is my impression that scientific explanation in general is no longer at the center of ongoing debates.[67] Most people are aware of the basic positions and their respective problems, but serious alternatives or radically new proposals are lacking. Present discussions focus on more specialized topics, such as the extent to which all explanation is contrastive, levels of explanation, explanation in the special sciences, idealization and abstraction, etc. Nonetheless, of course, the theories presented in this chapter serve as the background without which it is hard to make sense of these discussions.[68]

[67] As an exception to this observation, I should mention recent new work in the causal tradition by Woodward (2003) and Strevens (2004a; 2004b).

[68] Thanks to Erik Weber for commenting on an earlier version of this chapter.

CHAPTER VI

# Explaining the Behavior of Artifacts

## 6.1 Introduction

In this chapter I look at explanations that account for an artifact's behavior on the basis of its physical structure. For lack of a generally accepted or better term, I will call them *artifact explanations*.

Before we start, let us pause for a moment to get the big picture back in view. You will recall that the central topic is the relation between the intentional and physical perspective on artifacts as exemplified in the relation between an artifact's function and its structure. In chapter 2, I distinguished three aspects of this function-structure relation. A description of an artifact's function differs from a description of its structure in three respects: (a) It is a dispositional as opposed to a categorical description, i.e., a description in terms of behavior; (b) it is often fairly abstract as opposed to concrete and detailed; and (c) it presupposes intentionality, while a description of its physical structure is non-intentional. To get from function to structure, or vice versa, one has to deal with these three aspects. I also argued that, in the context of explanation, this is done in two parts. One part consists of artifact explanations that account for an artifact's abstractly described behavior, ultimately in terms of a concrete and categorical description of its physical structure. The other part then combines these explanations with an adequate function theory, which identifies the intended behavior that is the artifact's function. In explaining this intended behavior artifact explanations are drawn into the intentional realm, so to speak. They become explanations of how an artifact can fulfill its intended behavior, or of how it is or was supposed to do so. I talk about that in the next chapter.

The account of artifact explanation that I will develop fits in the causal-mechanistic explanation. The reasons for this are twofold. First and most importantly, this account fits actual examples of such explanations best. We will see this below. Second, the choice for this account of explanation is almost inevitable because of its intellectual origins. The general idea of mechanistic explanation has its roots in the mechanistic natural philosophy of the 17th century. The very core of this tradition was that nature must be understood and explained by considering it as a huge collection of mechanisms, as is testified by the following quotes.

> I was greatly helped by considering human artefacts – and I should say that the only distinction I recognise between artefacts and natural bodies is that, for the most part, the functions of artefacts are carried out by mechanisms big enough to be easily perceived by sensations. ... Certainly there are no explanations in mechanics which do not apply also to physics, since physics is a part or species of mechanics. (Descartes 1999: IV §203)[69]

> In the true philosophy, in which one considers the cause of all natural effects in terms of mechanical motions. This, in my opinion, we must certainly do, or else renounce all hopes of ever comprehending anything in physics. (Huygens 1945: 3)[70]

> It seems to me that the test of 'Do we or not understand a particular subject in physics?' is, 'Can we make a mechanical model of it?' (William Thomson, Lord Kelvin, 1884, cited in Kargon and Achinstein 1987)

Of course later proponents of the causal-mechanical model of explanation abandoned such metaphysical commitments, but it remains true that mechanistic explanations were modeled after artifact explanations. Hence, the claim that technical artifacts are susceptible to mechanistic explanation borders on being tautological.

In developing an account of how an artifact's behavior is explained the tradition of mechanistic explanation is a natural if not inevitable place to start. (I do not mean to imply that the causal-mechanical model is the best account of explanation in general; I leave it at the modest claim that it is an adequate account of artifact explanations.) As we saw in the previous chapter, the key tenet of the causal-mechanical account of explanation is that phenomena are explained by unearthing how they are brought about by an underlying mechanism, i.e., a system of spatiotemporally arranged causally interacting entities, which together produce the explanandum phenomenon. For natural phenomena, this raises epistemic accessibility problems. We do not always know how to cut nature at its joints and reveal a system's proper parts, nor do we always have the means to experiment with or manipulate any natural system. Furthermore, it has been argued that some systems give rise to holistic or emergent phenomena, which cannot be accounted for in terms of the system's parts. Some systems are said to be indecomposable or virtually so, which means that their components and their behaviors are so tightly interwoven that it becomes very hard to understand the whole system by studying its parts (Simon

[69] See w*ww.philosophy.leeds.ac.uk*/GMR/*hmp/homepage.html*.

[70] Available through *www.gutenberg.org/dirs/1/4/7/2/14725*.

1996; Bechtel and Richardson 1993; Wimsatt 1994). Such problems disappear in the case of technical artifacts.[71] After all, we ourselves design and make artifacts. We decide what their components are and what these components are supposed to do. Artifacts thus form a special case for mechanistic explanation, one in which we have perfect knowledge of the mechanism, its parts, their material constitution, and, possibly, their behaviors.[72] This makes artifacts particularly felicitous objects to study from the perspective of mechanistic explanation.

This chapter thus serves a double purpose. It provides us with an account of artifact explanation, which is a direct contribution to the project of this thesis. In addition it offers a clear view on mechanistic explanation since artifacts have none of the aforementioned epistemic accessibility problems. This may help to further clarify and develop this model of explanation.

The next section starts with some examples of artifact explanations to show that designers indeed give the sort of explanations that my philosophical account of artifacts says they do. I analyze the examples in section 6.3. In this first-pass analysis I distinguish two complementary explanatory strategies. In further fleshing them out, I argue that both strategies deliver self-contained explanations in section 6.4. Section 6.5 deals with some objections and caveats. After that, the primary project of the chapter is completed, but it turns out that this project has a noteworthy corollary in that it reveals a number of blind spots in recent treatments of mechanistic explanation. I review this work and point out some oversights in it in section 6.6. The conclusion in section 6.7 gives a recap of the main findings.

## 6.2 Examples of Artifact Explanations

A few examples should give the reader an impression of the type of explanations we are dealing with in this chapter. I will not try the reader's technical sophistication beyond necessity by elaborating the workings of discouragingly complex artifacts. Anyone should be able to grasp the workings of a paperclip, Edison's phonograph, and an early computer mouse. Fortunately, nothing essential for the present purposes is lost by looking at such relatively straightforward examples. The explanatory principles that I am interested in apply just as much to highly complicated machinery as they do these simple

[71] For the sake of simplicity I exclude artifacts with emergent properties (if such there be). For them, there would be similar epistemic accessibility problems.

[72] Note that we have perfect knowledge of the material constitution of the artifact and its components, because we consciously choose components and materials. We do not necessarily have perfect knowledge of all of an artifact's behaviors however, since our understanding of what things can do may be incomplete, flawed, or both.

artifacts. It is a coincidence that all three examples stem from patents; one finds similar explanations in engineering handbooks.

### *6.2.1 Paperclip*

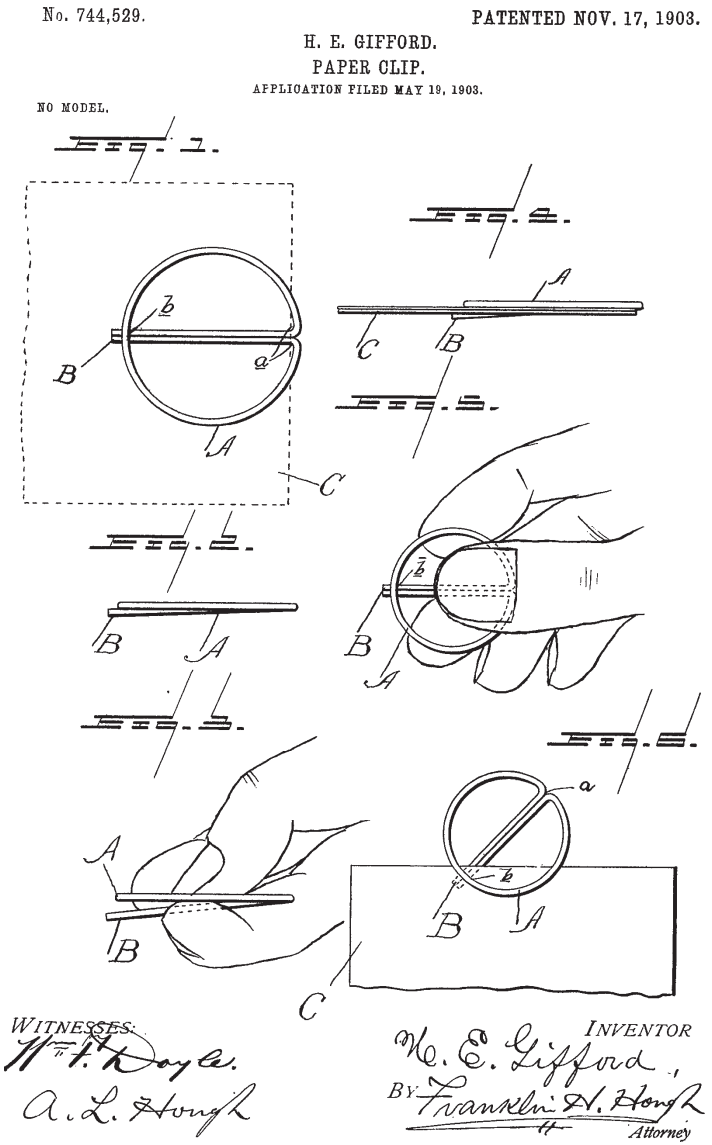
My first example is really straightforward; it is a paperclip from 1903. I found it in one of Henry Petroski's wonderfully entertaining books, *Invention by Design* (1996). As to the explanation of how this paperclip works, Figure 5 says most of it but let me add a few things in words. A paperclip does not just slide on papers by itself; we have to put it in place by opening the clip with our fingers and positioning it, as the patent pictures show vividly. What subsequently makes the paperclip do its job is springiness; or rather the force exerted on the papers that results from the paperclip's spring action. Like any material, the metal of which the paperclip is made has a characteristic springiness or elasticity. Its elastic behavior is described by Hooke's law: 'As the extension so the force.' More prosaically: $F = -kX$, where $F$ denotes force, $X$ extension, and $k$ an object's elasticity constant. Up to a material's elastic limit, the more we stretch something, the more resistance it offers to further stretching. Beyond this limit permanent deformation occurs. This is the key to paperclip behavior: flexing the loops out of their normal plane position causes them to exert a force in the opposite direction, and thus they hold the sheets of paper together.

Let us also listen to Gifford himself explaining some specifics of his paperclip. After giving a detailed description of its material constitution, shape, and size, he accounts for a few noteworthy features of his invention (all quotes are from the patent specification).

> The size and circular form of the clip will serve to provide a uniform pressure extending over a considerable surface of paper and that the paper engaged by the clip will be securely held in place without defacement or injury, and it will also be noted that the paper may be readily removed when desired from the clip without danger of tearing.

> The essential features of my form of clip reside, first, in the circular body portion of the clip through which the thumb passes in pressing upon the free parallel ends of the wire, this circular form also affording a bearing for the fingers pressing the opposite direction, thereby permitting the clip to be opened readily, and, second, in the distance between the spring-shoulder *a* and the point of contact *b*, where the free end portions *B* of the wires press against each other. By this construction the spring action of the cross-wires is not weakened in use, and the clip can in consequence be used for an indefinite time without injury.

***Figure 5*** *H.E. Gifford's 1903 Paperclip (Patent No. 744,529)*[73]




[73] This picture and the following ones are available through the website of the United States Patent and Trademark Office, *www.uspto.gov*.

> It is not necessary to bend the papers in passing them between the wires of the clip, while the size of the body portion is such as to prevent wrinkling or disfigurement of the papers, as above stated, and it will not be necessary that the edges of the papers be arranged uniformly in order that the same should be firmly gripped by the arms of the clip.

I interpret these quotes as showing that the function of Gifford's paperclip comprises much more than simply holding together a few papers; it is to do so without 'defacing or injuring the paper,' to do so repeatedly 'for indefinite time,' and to do so even if papers are not 'aranged uniformly.' Moreover, the paperclip can be 'opened readily,' 'removed readily without danger of tearing,' and prevents 'wrinkling or disfigurement of the papers.' All of these loosely described dispositions are accounted for by the shape, size, and material constitution of the paperclip, i.e., by its properties categorically described.

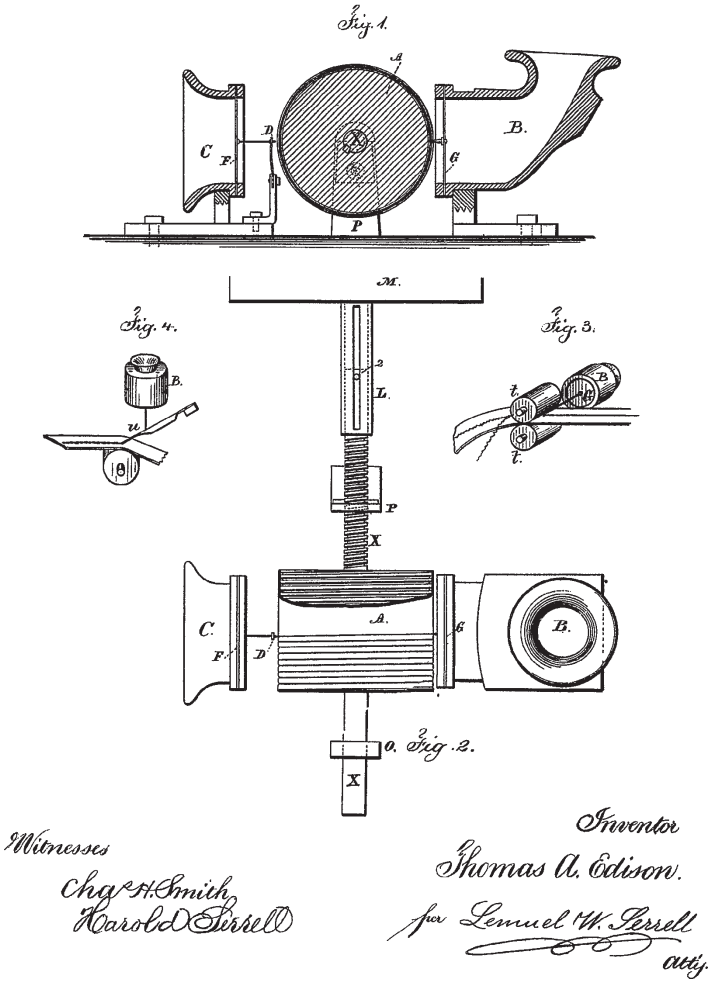
### *6.2.2 Phonograph or Speaking Machine*

Thomas Edison's 'phonograph or speaking machine' is without a doubt a more sophisticated device than the paperclip, but its operational principles are still relatively easy to grasp. The patent drawings are shown in Figure 6. The upper drawing, labeled 'Fig. 1', is a cross sectional view of the phonograph, and the lower ('Fig. 2') gives a plan view of it. The smaller drawings to the left and right ('Fig. 3' and 'Fig. 4') show alternative recording devices that may be used instead of the one in Fig. 1 and 2; we ignore them here. The key discovery leading to this device is

> that a diaphragm or other body capable of being set in motion by the human voice does not give, except in rare instances, superimposed vibrations, as has heretofore been supposed, but that each vibration is separate and distinct, and therefore it becomes possible to record and reproduce the sounds of the human voice.

The capital letters in the drawing indicate the principal components of the phonograph. Edison describes them in considerable detail; I will summarize his points. Component *A* is a cylinder with a helix-shaped groove pattern on its surface. Recording material (preferably metallic foil, according to Edison) is fastened on the surface of *A*. Cylinder *A* is secured to shaft *X* (note that *X* has screwing thread on one end), which is supported by *P* and *O*. Support *P* also has screwing thread on its inside. At one end (the upside of Fig. 2, the end with the screwing thread), *X* enters tube *L*, which has an elongated slot in it. *L* can be rotated by device *M* and because *X* has a pin (*2*) secured to it that is sticking through the slot in *L*, *X* can rotate as well. With the action of the screwing thread

***Figure 6*** *Thomas Edison's Phonograph (Patent No. 200,521)*[74]




[74] In addition to the USPTO website, this patent is also available through the excellent Thomas Edison archive at Rutgers University, *edison.rutgers.edu*.

on *P*'s inside, *X* will start moving towards *O* (which is why the slot in *L* has to be elongated). Component *B* is a mouthpiece and *G* is its diaphragm, which has a hard pin-like indenting point at its center poised exactly perpendicular to the surface of *A*. In its operating position, this point indents the recording material slightly. On the other side, a similar assembly is placed: *C* is roughly the same sort of device as *B* but functions as a speaker. Its diaphragm *F* also has an indenting point fixed at its center, but this one is held in place by a light spring *D* that presses it in the grooves made in the recording material on *A*'s surface. Note that *B* and *C* will not be activated at the same time (since *B* is for recording and *C* for playing sounds).

If someone speaks into the mouthpiece *B* the air vibrations carrying the sound will cause the diaphragm to vibrate so that the indenting point at its center will make indentations in the rotating recording material on *A*'s surface. As *X*'s rotation makes *A* move in the direction of *O*, the recorded track becomes spiral shaped. The track's pattern of indentations will represent voice movements and strength. Recording can continue until *A* reaches its utmost position at *O*. The recording material can then be removed from *A* and stored for later use, or it can be 'rewound' to its initial position to be played again. I quote Edison again on the replaying of a record.

> This spring and point [*D*, JdR] are so arranged as to fall exactly into the path of all the indentations. This spring is connected to the diaphragm *F* of *C* by a thread or other substance capable of conveying the movements of *D*. Now, when the cylinder is allowed to rotate, the spring *D* is set in motion by each indentation corresponding to its depth and length. This motion is conveyed to the diaphragm either by vibrations through a thread or directly by connecting the spring to the diaphragm *F*, and these motions being due to the indentations, which are an exact record of every movement of the first diaphragm, the voice of the speaker is reproduced exactly and clearly.

Towards the end of the description, Edison also comments:

> It is obvious that many forms of mechanism may be used to give motion to the material to be indented. For instance, a revolving plate may have a volute spiral cut both on its upper and lower surfaces, on the top of which the foil or indenting material is laid and secured in a proper manner. ... A wide continuous roll of material may be used, the diaphragmic devices being reciprocated by proper mechanical devices backward and forward over the roll as it passes forward; or a narrow strip like that in a Morse

> register may be moved in contact with the indenting-point, and from this the sounds may be reproduced.

We see a similar style of explanation as above. The overall behavior of the phonograph is decomposed into a number of sub-behaviors, of which the various components take care. Their ability to do so is explained by where they are located relative to one another and some of their own specific properties, which are often described categorically (shapes, textures, material, etc.). The last quote is interesting because it shows how Edison recognizes the possibility of functional equivalents; underlying mechanisms that take care of the same behavior in a slightly different manner. The recording-by-indenting behavior can be implemented by different structures, but that is relatively unimportant for the explanation of how the phonograph works.

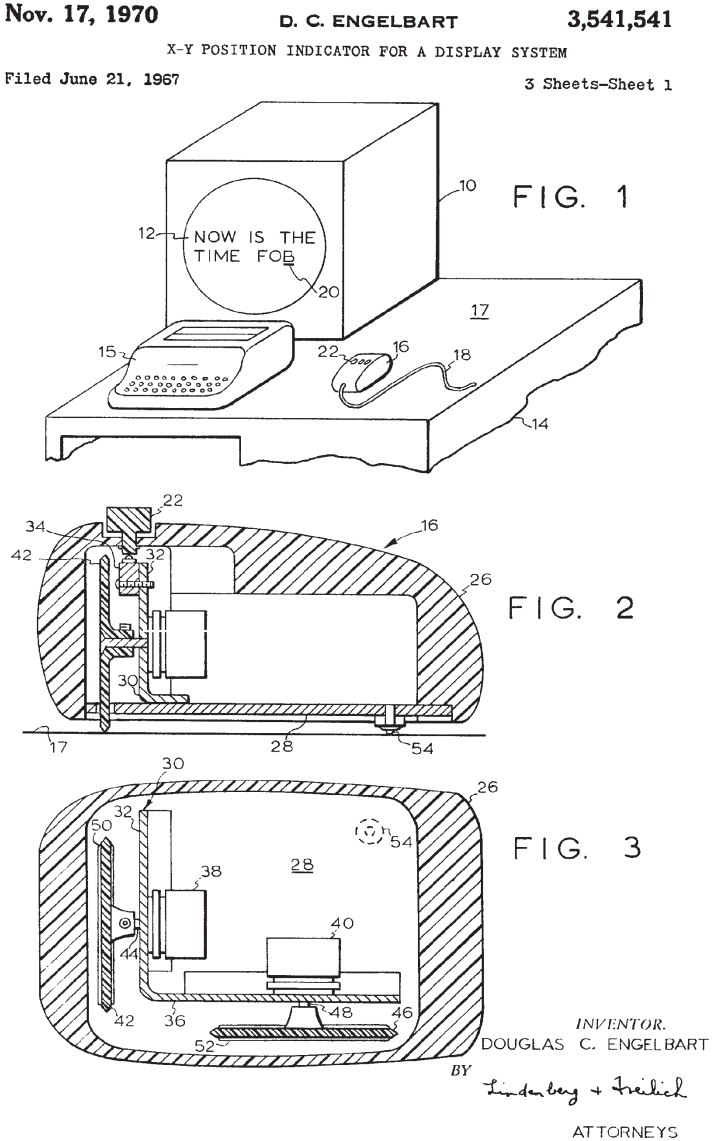
### *6.2.3 X-Y Position Indicator*

The third and final example is one of the earliest computer mice from 1970. Its inventor, Douglas C. Engelbart, calls it an 'X-Y Position Indicator for a Display System.' Figure *7* shows the patent drawings. Most of us will be familiar with the operation of a computer mouse so I will keep the explanation brief. In contrast to modern mice, this one has two large wheels (*42* and *46* on the drawings) to register the movements over a surface. It rests on these two wheels and a ball bearing, *54*, to provide stability and yet allow smooth movement. The wheels are connected to potentiometers that change their electric resistance as the wheels rotate. By measuring resistance, a signal is generated that encodes the indicator's movements in the *X* and *Y* direction.

> When the position indicator control is moved over the cabinet *17*, or any other surface, the *X* and *Y* position wheels rotate. Inasmuch as the *X* and *Y* position wheels *42* and *46* are mounted on axes that are perpendicular to each other, the *X* position wheel *42* rotates by an amount equal to the movement in one direction which may be defined as the *X* direction, while the *Y* position wheel *46* rotates an amount equal to the movement in a perpendicular or *Y* direction. As the wheels move, the shafts of their respective potentiometers rotate, and the resistance of the potentiometers enable continuous measurement of the resistance, and therefore of the *X* and *Y* positions of the indicator control *16*.

Engelbart has also given some thought to possible functional equivalents for the potentiometers. Other mechanisms are available to generate signals that track the rotation of the wheels. He describes two of them at length in his patent.

***Figure 7*** *Engelbart's X-Y Position Indicator (Patent No. 3,541,541)*



> While a potentiometer may be connected to each of the two wheels on the indicator control, other devices can be used for generating signals indicating rotation of the wheels. One such device is a shaft position encoder which produces digita [sic] output corresponding to the angular position of the wheel. ... Still another means for indicating position of a wheel is an incremental encoder and counter. An incremental encoder generates an up indicating pulse each time the shaft moves by a certain increment of rotation in one direction and generates a down indicating pulse when the shaft moves in the other direction. These pulses are transmitted to an up-down counter, which provides a digital output equal to the sums of the up inputs minus the sum of the down inputs.

Finally, two more quotes that contain explanations of some further specific dispositions of the X-Y position indicator, again by simple categorically described properties such as relative locations, shape, and size.

> The use of only three points of contact, comprising two wheels and the ball bearing support, help to assure that both wheels will constantly remain in firm contact with the surface on which the position indicator control rests.

> The use of an indicator control which rests firmly on a surface enables the operator to accurately maintain position with a minimum of muscle effort, since the indicator control remains stationary unless some force is applied to it. The use of relatively large wheels having appreciable, even if small, moments of inertia, reduces jittering of the indicator control and promotes smooth movement.

## 6.3 Preliminary Observations and Two Explanatory Strategies

Let us now make a first attempt at clarifying what is going on in these examples. I will offer some observations on both the explanandum and the explanans.

### *6.3.1 The Thing Explained*

First a brief observation on what is *not* happening in the examples to justify my contention that the connection between function and structure in the context of explanation splits in two parts. Nowhere do we observe attempts to explain or justify that the artifact in question has a particular behavior as its function. Edison neither justifies that his phonograph has the function to record and reproduce sounds, nor does he argue for the claim that recording sound, rather than, say, decorating the living room, is the phonograph's technical function. That recording and reproducing sounds is the phonograph's technical function

is simply presupposed as unproblematic and obvious throughout. What he does explain is how his phonograph works, or how it is that the device has the right dispositions to record and reproduce sounds. The same goes for the other examples. This might seem like a slightly trivial observation, but I include it nonetheless to make the point that analyzing artifact explanation should not be conflated with developing an adequate theory of artifact functions. The two are separate projects and they will not benefit from being run together. In the next chapter we will have ample space to develop this point.

On a more positive note, what is explained in these examples is *how the artifact works*, i.e., how it is able, in virtue of its physicochemical makeup, to exhibit specific behaviors, namely those its designer intended it to have; the ones that are its function. In other words, the explanations show how the artifact has the dispositions that enable it to function correctly. The question which of an artifact's many possible behaviors is its function is largely immaterial to this explanandum. One could just as well explain any old behavior of the artifact using the same style of explanation. It is just that, by most standards, the behavior with which the artifact fulfills its function naturally stands out as the behavior in need of explanation. Typically, an artifact's function is the very reason for its existence and the behavior that implements that function is what typically makes the artifact useful and interesting for us. Therefore it stands to reason that this behavior is normally the one that must be explained.

More specifically, the explanandum of these explanations is an artifact's having a particular disposition, a capability to exhibit certain behaviors when manipulated in certain ways in certain circumstances. The canonical form of such explananda is something like 'the artifact $\phi$'s when it is $\chi$-ed' or 'artifacts of this kind $\phi$ when they are $\chi$-ed.' A paperclip holds papers together when it is slid onto them; the phonograph records sounds when it is used in the way described by Edison; the computer mouse moves a cursor around on a screen when it is moved across a surface; etc. Note again that, although the behavior being explained will often be the artifact's function, the explanandum is not 'the technical function of this artifact is to $\phi$ when $\chi$-ed.' That this particular behavior is the artifact's function is simply taken for granted in the examples. Further, the explanandum says nothing about the underlying mechanism, only that it produces output of a certain type when it is manipulated in an appropriate manner in suitable circumstances. So what needs to be explained is a dispositionally described property of an object. As we saw in the previous chapter, mechanistic explanation is well suited to provide such explanations.

Consider briefly what happens when we want to explain how a malfunctioning artifact can fulfill the behavior that is its function. We produce an explanation along the lines of the above examples. This explanation must then contain at least one falsehood since it (allegedly) shows how the artifact

can show its intended behavior. But *ex hypothesi* it cannot do that. That means that at least one of the disposition ascriptions to the artifact or its components must be wrong. At the same time, it seems that the original explanation is still perfectly in order, for it tells you – truthfully – how the artifact in question is *supposed* to work. This explanation recounts how the artifact's designer believes or believed it to work and thus ascribes the behaviors and sub-behaviors to the artifact and its components that the designer *intended* them to have. If the artifact malfunctions, one or more of these intended dispositions are not among its actual dispositions. A fuller elaboration of this point must wait. It suffices for now to note that artifact explanations can inherit the intentionality of function ascriptions in the way just indicated and to be specified in the next chapter.

### *6.3.2 The Thing Explaining It*

Let us now look at the explanans of an artifact explanation. A first-pass characterization of what is going on in the examples is that the explanans of an artifact explanation identifies different sub-behaviors of the overall behavior, shows their spatial configuration, and describes their interactions. In other words, it looks as if an important part of the explanans is a functional decomposition. We saw, for instance, how the overall behavior of the paperclip turned out to have several sub-behaviors, which Gifford mentions and discusses in turn. Edison describes the behaviors of the mouthpiece on the recording side of his phonograph, then that of the cylinder with the indenting material on it, and then that of the replaying side of the phonograph. He also describes the behaviors of the diaphragm, the indenting point, and the metallic foil separately, and so on. The same happens with the computer mouse: The behavior of the wheels is recounted, that of the ball bearing, the mouse buttons, the wire, etc.

The explanans further describes the artifact's components and their physicochemical make-up and says that these components take care of all behaviors and dynamic interactions. At various points we also get information about how it is that the components themselves have the dispositions they have. It is explicated how certain behaviors are made possible by underlying mechanisms. The explanations seem to end in relatively low-level and concrete categorical descriptions of spatial, physical, and chemical properties of the artifact and its components. Examples of this, too, abound. Gifford explains the sub-behaviors of his paperclip by reference to its circular shape, its size, and the location of the free ends of the wire relative to the body of the clip. The very discovery that led to Edison's phonograph is that a diaphragm, i.e., a certain kind of artifact, the categorical properties of which will have been implicitly understood by the reader, does not produce superimposed vibrations when it is set in motion by the human voice. The spatial configuration of the cylinder and

how it is attached to an electromotor or similar device explain the movement of the indenting material, while the helical pattern on that material explains the trajectory of the indenting point, etc. The mouse patent shows similar bits of explanation: The mouse remains in form contact with the surface on which it rests because it has three contact points, moving it around goes smoothly because of the size of its wheels, the wheels encode movement in the *X* and *Y* direction because they are perpendicular to one another, and so forth.

These two sets of examples embody two different ways to explain an artifact's overall behavior, which are apparently often used in conjunction. By way of a first rough approximation, these strategies can be characterized as follows. The first explains the overall behavior by decomposing it into sub-behaviors, so making clear how the complex overall behavior of the artifact is realized by various simpler or more basic sub-behaviors. The second adds to this how the various sub-behaviors are explained by the categorically described spatial and physicochemical properties of the artifact and its components.

I think the use of functional equivalents in the explanations above lends support to the idea that these are indeed two different strategies. We saw how Edison says that 'it is obvious that many forms of mechanism may be used to give motion to the material to be indented.' I did not quote this before, but he also says about the mouthpiece *B*: '[It] may be of any desired character, so long as proper slots or holes are provided to re-enforce the hissing consonants.' In the third example Engelbart says, referring to the potentiometers in his mouse, that 'other devices can be used for generating signals indicating rotation of the wheels.' Apparently, then, the exact underlying mechanism taking care of a behavior is not important for understanding the overall working of the device. As long as something has the disposition to 're-enforce hissing consonants,' its further physicochemical make-up are largely irrelevant. Similarly, it does not matter what the exact mechanism is, as long as it generates some sort of signal that encodes the movements of the wheels. Such points show that designers do not find it necessary to always explain the dispositions of the artifact's components on the basis of their categorically described properties. To explain how an artifact works, it often suffices to show how its overall behavior is made possible by (various levels of) lower-level simpler sub-behaviors.

### *6.3.3 Two Explanatory Strategies*

These examples illustrate what I take to be two complementary and partly overlapping yet different explanatory strategies for explaining artifact behavior. The first strategy focuses purely on behaviors of the artifact and its components and points out (by mere ostention) which components take care of which behaviors. It just mentions the components but treats them as black boxes with behaviors, remaining silent about their physical constitution in

terms of categorical properties. By doing so, it explicates how the complex overall behavior of the artifact is realized by increasingly simpler sub-behaviors. In the end, it provides a spatiotemporal breakdown of the overall behavior into several levels of sub-behaviors, which shows how a complex behavior is implemented by a collection of simpler behaviors; its behavioral parts. By way of analogy, think of algorithms. Algorithms perform complex tasks by breaking them down into a number of simpler tasks that, when executed in the right order, accomplish the complex task. In the same vein, this explanatory strategy shows how the behavior of an artifact is realized by various simpler interacting sub-behaviors that are manifested in an appropriate order and at the right places.

In effect, we can now see, this strategy gives the functional decomposition, or a part of it, of the intended behavior of the artifact in question. A functional decomposition, as we have seen in chapter 4, is a breakdown of the artifact's overall intended into possibly many nested levels of sub-behaviors. The information it conveys is how the overall behavior of the artifact is made possible or realized by various simpler or more basic behaviors. And that is indeed an answer to the sort of explanatory questions we are considering here.

I argued at length in chapter 4 that functional decomposition has different senses. It is interesting to note that these different senses reappear here. I already gave examples of the straightforward sense of merelogical decomposition of behavior.[75] The seemingly simple behavior of the paperclip turned out to have several component-behaviors, which Gifford explains one by one. The phonograph's behavior obviously splits into its recording and replaying behavior and these two behaviors in turn have further finer-grained parts. In its supervenience senses, functional decomposition makes abstract behaviors more concrete. The above examples also provide ample illustrations of that. Of course in the obvious sense that the ways in which the three artifacts explained are all concretizations of their abstractly described behaviors, but there are also more specific examples. The functional equivalents I discussed before are also good examples. The behavior 'moving the material to be indented' is made more concrete by the different mechanisms Edison describes. Just as the abstract behavior of generating a signal that encodes the movements of the mouse is made concrete by what a potentiometer does, what a shaft position encoder does, and what an incremental encoder and counter do. Or, finally, consider this quote from Engelbart's patent: 'Three buttons *22* are located on the indicator control *16* for operating switches within the indicator control to allow currents to flow through conductors of the wire *18*. The switches

[75] Recall that these examples may also be read in the member-of-set sense, depending on one's taste in event ontology.

may be used to cause changes in particular areas of the display, or for other purposes.' The abstract behavior 'causing changes in the display' is realized by the concrete behaviors of buttons, switches, and currents flowing through a wire.

The second strategy opens up the dispositional black boxes and explicates the categorical bases for the properties. It explains how the spatial and categorical physicochemical properties of the artifact and its components enable them to show the behaviors they show. This strategy starts from the physical components of the artifact, describes their relevant spatial and categorical physicochemical properties and shows how these properties give rise to the relevant dispositions. For the rest, it overlaps the first strategy, because it must also describe how these behaviors, in their specific spatiotemporal configuration, result in the overall artifact behavior. Without that, it would not provide a complete answer to the explanatory question. To use the algorithm analogy once more, we can think of the second strategy as explaining how the material realization of the algorithm is able to perform the required steps. For instance, by showing how the algorithm is hard-wired in an electronic device with the right dispositions, or how appropriate lines of software code realize the required computations or logical operations.

Before I look at the two strategies in more detail and point out some of their features, let me label them and characterize them schematically. The explanandum for both strategies is an artifact's (*a*'s) having the ability to show a particular behavior, *b*; typically that associated with its function. The two strategies are two ways to generate the explanans. Since the first strategy starts from the overall behavior of the artifact and works its way 'down' to lower-level component-behaviors and less abstract behaviors, I dub it the *Top-Down Strategy* (TD). The second strategy starts from lower-level physicochemical properties (both categorical and dispositional) of the physical components of the artifact and then works its way upward to the overall behavior; I therefore call it the *Bottom-Up Strategy* (BU).[76] Here are brief characterizations of both strategies.

> (TD) Describe *b*'s sub-behaviors $b_1$, $b_2$, ..., $b_n$, their sub-behaviors, and so on. Continue until the level of basic sub-behaviors. Include the spatiotemporal configuration of the behaviors. (Put succinctly: Give the functional decomposition of *b*.) Point out ostentatively which structural components take care of the basic sub-behaviors.

[76] By adopting this terminology I do not mean to endorse the strictly hierarchical picture of nature and the place of categorical and dispositional properties in it that I rejected in chapter 4.

> (BU) Describe *a*'s physical components $c_1$, $c_2$, ..., $c_n$, sub-components, and so on. Continue until the level of basic physical components. To the extent that this is possible, give the categorical bases for all of the $c_i$'s relevant dispositions. Then, also give the functional decomposition to show how the lower-level basic behaviors just accounted for compose into *a*'s overall behavior *b*.

A few clarifications. The sense of 'possible' in (BU) is metaphysical. I want to leave it open whether all dispositionally described properties can be explained by categorically described properties. Some philosophers have argued that, necessarily, all dispositions have a categorical basis (e.g., Armstrong 1968), but others think that some dispositions are ungrounded (e.g., Mumford 2006) and that properties can be dispositional 'all the way down.' I do not want to build either position into my explanatory strategy, although it does seem plausible that the dispositions of the artifacts I am considering in this thesis will typically have a categorical basis. The only serious candidates for ungrounded dispositions seem to be those of subatomic particles investigated by fundamental physics.

Second, these characterizations are idealized in that they stipulate what complete (TD) or (BU) explanations would look like. For various sort of pragmatic reasons, it is often perfectly all right to give partial explanations of either the top-down or the bottom-up ilk. Someone may be interested in just one or a few component(s) of the artifact, people may have qualitatively and quantitatively different background knowledge, or other pragmatic factors may make it superfluous to go into each and every detail. As I said at the beginning of chapter 5, this phenomenon is common to every explanatory activity and I will disregard it in what follows.

Readers who are acquainted with the canons of mechanistic explanation, will have noticed how these strategies resemble Bechtel and Richardson's (1993) mechanistic explanatory strategies:

> The [analytic strategy] is to isolate components physically within the system and then determine what each does (the goal is to use the knowledge of components to reconstruct how the system as a whole operates). The [synthetic] strategy is to conjecture how the behavior of the system might be performed by a set of component operations, and then to identify components within the system responsible for the several subtasks (ibid.: 18).

Rougly, my (TD) strategy is their synthetic strategy and my (BU) strategy is the analytic strategy. The basic idea seems to be the same, but there are two

important differences. First, Bechtel and Richardson's strategies do not depend explicitly on the distinction between categorical and dispositional descriptions. In the quoted characterization it even looks as if their strategies do not provide categorical underpinning of behaviors at all, but the rest of the book makes it clear that this can be part of either strategy. Second, the analytic and the synthetic strategy should yield the exact same result for artifacts. An artifact's functional decomposition objectively determines what the 'set of component operations' is; there is no need to conjecture what this set might be. Furthermore, all the artifact's physical components are created, selected, and put in place to take care of one or more of these sub-behaviors. As a result, it does not matter where you start; at the 'bottom' with physical components or at the 'top' with overall behavior. The result should be the same in both cases. Applied to artifacts, then, Bechtel and Richardson's strategies turn out to be merely pragmatically different ways of constructing the same explanans, whereas my strategies produce two different (albeit partially overlapping) explanantia. Whether this is also the case for natural (living) objects is an open question; the answer depends on whether there are objectively correct physical and functional decompositions of them and their behavior respectively (and, if so, whether we can ever know them). Without objectively correct *functional* decompositions the synthetic strategy may lead to very different sets of component operations, because behaviors are always multiply realizable by sub-behaviors. Without objectively correct *physical* decompositions the analytic strategy may carve up objects in different ways and subsequently also identify different behaviors.[77] So for natural objects, Bechtel and Richardson's strategies may indeed lead to different results, which, in the absence of an objectively correct explanans or the epistemic means to distinguish the correct explanans from others, may all shed valuable light on the object's workings.

Another related classic source is of course Robert Cummins (1975), who distinguishes between the (1) Subsumption Strategy and (2) Analytical Strategy for explaining dispositions (ibid.: 758–760). The former explains an object's disposition by subsuming the dispositional regularity under a general law, which holds for a certain general class of objects and not just objects having this particular disposition. The latter explains a disposition by decomposing it into sub-dispositions of the object or its components. Spatiotemporally ordered

[77] Artifacts with unknown functions as are sometimes found in archeological excavations and artifacts in situations of reverse engineering are somewhat similar to natural objects in these respects. We do not know what their functional decomposition is, although in virtue of their being intentionally designed they do have one. The physical decomposition of artifacts will normally be fairly clear because physical parts tend to be made of different materials, fixed to one another by means of clearly distinguishable securing means, etc.

manifestation of these sub-dispositions should amount to manifestation of the overall disposition.

Clearly, Cummins's Analytical Strategy is very similar to my (TD) strategy. He even explicitly says his strategy is a form of functional analysis and this term is a synonym for functional decomposition. I would be happy to consider my strategy as a somewhat more developed and precise version of the Analytical Strategy. The Subsumption Strategy, however, differs from my (BU) strategy. It does not explicitly stress the explanatory value of categorical descriptions of properties, even though the general laws may be laws that relate an object's categorical properties to its dispositions. In my opinion, the seemingly uncritical reliance of this strategy on the D-N model of explanation and an unanalyzed notion of general law renders it rather problematic. A final difference seems to be that my two strategies each produce self-contained explanations. Cummins's strategies should apparently be used in conjunction; the Subsumption Strategy takes over when Analytical Strategy stops.

### *6.3.4 Features and Refinements of the Two Strategies*

To further explicate the two strategies, I want to draw attention to three of their features. The first is best understood in connection with the remark I made above about the strategies being somewhat idealized. There is another source of context relativity that is explicitly built into the strategies. It is reflected in the phrase 'basic' for both sub-behaviors and components. That term should not be understood in any deep metaphysical sense, but as a context-relative qualification for behaviors or components that are considered not to be in need of further decomposition in the given context. The context will primarily consist of the particular engineering sub-discipline, but may also include factors related to the design of the artifact being explained. It is the same sense of basicality that we came across in chapter 4 when we were discussing the stopping point for processes of functional decomposition. For instance, some electrical engineers will take microchips as basic components for, say, the computers they design while others are concerned with improving that microchip. For the former, the microchip and its behaviors will count as basic, but for the latter the same chip and behaviors are the very opposite of basic. For them, presumably, the materials of which the chips are made and their dispositions will be basic. For material engineers, even the material itself maybe non-basic and they may think about behaviors of molecules or atoms as properly basic.

A consequence of this point is that the two strategies can be combined into an alternating sequence of (TD) and (BU) explanations. For an explanation of a computer's workings the behavior of the microchip and the chip qua physical component may be considered as basic and the additional explanatory

information provided by the (BU) explanation of the chip's behavior may consist of no more than a statement of the precise type of microchip it is (which qualifies as a categorical description of it because kind names of things and materials are in principle categorical descriptions of it).[78] The microchip's behavior may subsequently become the explanandum of another (TD) explanation, which does decompose the chip's behavior, as well as its physical parts. The (BU) explanation of the chip's components' dispositions could then consist of citing molecular structures of the materials of which the chip is made. This context relativity of basicality – and hence the dividing line between (TD) and (BU) explanations – is a direct consequence of the relativity of the dispositional-categorical distinction that I discussed in chapter 4. Since the distinction between the (TD) and (BU) strategies hinges on that between the dispositional and the categorical, any context relativity clinging to the latter distinction automatically carries over to the former.

The second feature should not come as a surprise, given what I have said about functional decomposition before. In the above discussion of the explananda of artifact explanations, I noted that they are phrased in purely dispositional terms. This observation carries over to the (TD) strategy; it, too, uses only dispositional descriptions of the artifact and its components. It describes behaviors, not underlying physical implementations. That does not mean that the strategy is completely independent of the underlying physical details. The categorical details of the chosen implementation do matter but they are hidden under the surface of the first strategy, so to speak. As I argued in chapter 4, functional decompositions minimally contain categorical information about the spatiotemporal ordering of the behaviors described. And more often than not it is possible to infer additional categorical information by supplementing them with relevant background knowledge from the engineering practice. This dependence does not, however, alter the fact that there is no direct reference to the underlying implementation in the first strategy. Only in the final stage, when it is pointed out which physical components take care of various behaviors, do the physical components come into view. But even then no categorical descriptions of them are provided, for the pointing out is done by mere ostention. The (BU) strategy does give categorical descriptions; that is its defining characteristic and the key difference with the (TD) strategy.

Third, the two strategies employ two significantly different notions of decomposition. The first is the familiar notion of *physical*, or material, decomposition. It reveals the mereological make-up of artifacts and singles

[78] As we have noted before, many artifact kinds are named after the function artifacts of that kind typically fulfill. Arguably, such kind names do not qualify as categorical descriptions.

out the physical parts composing them. A physical decomposition of, say, a house identifies the component parts of the house such as walls, doors, windows, floors, ceilings, roof, and, on a lower level of detail, bricks, planks, doorknobs, glass panes, nuts, bolts, nails, screws, etc. The second notion is that of *functional* decomposition that we encountered in chapter 4 in its role as a tool for designing engineers. As we saw there, not only do artifacts have physical parts, they, or better: their intended behaviors, also have behavioral parts. The overall intended behavior of an artifact can be broken down into sub-behaviors that make up that overall behavior when executed in the right order at the right places.

It is possible, but neither necessary nor very typical, that there is a simple one-to-one correspondence between physical and functional decompositions so that each basic component $c_i$ of an artifact performs exactly one basic behavior, $b_i$. It is far more common, however, for some physical parts to take care of more than one behavior, and for some behavior-parts to be taken care of by more than one physical part. The (TD) strategy is almost exclusively about the functional decomposition; the (BU) strategy takes its starting point in the physical decomposition, but also needs the functional decomposition to form an adequate explanans.

This completes my outline of the two explanatory strategies. In the next sections I will be primarily concerned with the objection that there is no philosophically interesting distinction between my two strategies because the (TD) strategy merely produces a trickled-down version of the real explanation, which is generated by the (BU) strategy. Against this objection, I will maintain that both strategies produce self-contained explanations. The objection is based on misguided assumptions about what a good explanation is. So I argue in the next section.

### 6.4 Two Strategies, not One

A plausible reaction to my distinction between a (TD) and (BU) strategy might be to deny that they are in fact two interestingly different strategies. Someone may want to argue that a (TD) explanation is nothing more than an incomplete version of a (BU) one. Such an incomplete explanation may suffice for practical purposes in some circumstances, but at the end of the day it is not the full monty. A good – as opposed to merely satisfactory – explanation of how an artifact can fulfill its function should always end in the rock bottom of categorical physical properties. Reasonable as it may seem, I think this reaction is nonetheless mistaken. We should distinguish the two strategies not only for pragmatic reasons but also as a matter of principle. To argue this point I will show why and how (TD) explanations are perfectly self-contained and legitimate explanations. Because explanations of the (BU) variety contain all the

explanatory information one may wish for, I take it that *their* autonomy needs no defense. The suspicion is on (TD) explanations. To mount my argument, I need one background assumption and a more precise characterization of the sorts of information that my two strategies provide.

### *6.4.1 Preliminaries*

The assumption is a claim about the relative independence of ontology and explanation. Higher-level explanation has its own right, even if we accept a form of physicalism which says that reality is ultimately made up out of the elementary particles (or whatever stuff) that physics studies, and that all the 'real' efficacious causal interactions go on at that level.[79] It may be true that everything that exists is ultimately constituted by sub-atomic particles and in some sense reducible to these particles, but it does not follow from that that an ideal explanation must be couched in terms of these particles. It may also be true that only these basic particles engage in real causal interaction, but that, too, does not mean that every explanation, causal or other, ought to be expressed or even expressible in the vernacular describing these particles. This claim is not exactly new; there is a long-standing discussion in the philosophy of science on the desirability and feasibility of reductionism. It starts with the classic champions of reductionism like Kemeny and Oppenheim (1956), Oppenheim and Putnam (1958), and Nagel (1961). They were subsequently confronted with various anti-reductionist arguments (famously, e.g., Fodor (1974; 1997), Putnam (1975), but also Kitcher (1984), Kim (1992), Dupré (1993), and Cartwright (1999). While not all of these philosophers were directly concerned with explanation, their positions do have ramifications for a theory of explanation. After all, if a particular branch of science is irreducible to another, that is a good reason to allow for explanations in terms of both branches. Rehearsing this lengthy and continuing debate is entirely beyond the scope of this chapter. I will simply assume that, in general, higher-level explanation has its own right. In the appendix to this chapter, I do give three arguments that make this assumption plausible. Later in this section, I will offer specific arguments for the claim that explanations generated by my (TD) strategy are self-contained, good explanations.

Now I want to take a closer look at the sorts of information that the two strategies provide. Recall that the (TD) strategy gives the functional decomposition of the artifact's overall behavior and shows how programmed

[79] I do not mean to endorse this doctrine; it only serves as a worst-case scenario for my argument. It may be the case that the world contains several 'ontologically real' levels of organization and causality. In that case, however, the status of higher-level explanation will be much easier to defend than on the above doctrine.

manifestation of sub-behaviors, executed by the physical components, leads to the overall behavior. All this information is phrased in dispositional terms, black-boxing most details of the underlying physical realization (even though a functional decomposition does contain and suggest some categorical information). The (BU) strategy provides the categorical underpinning for all component behaviors. By explicating how the categorical properties of the artifact's physical components give rise to the relevant dispositions, this strategy complements the (TD) strategy. However, it also partly overlaps it by showing how the component behaviors add up to the artifact's total behavior.

The information provided by (BU) explanations is more *extensive* and more *specific*. In addition to all the information included in (TD) explanations, it also contains the additional story about the categorical bases for the dispositions of a particular artifact and its components. In connection with the familiar point about the multiple realizability of dispositions – the same behavior can be implemented by physically different structures – the (BU) strategy also provides more specific information. It tells you how one particular artifact (or type of artifact) realizes its component behaviors and so pinpoints one way to realize these behaviors, out of possibly many ways in which they could have been realized. Another way to make the same point is to say that the functional decomposition given in the (TD) strategy is more coarse-grained in that it does not distinguish between different material implementations of the same behavior. The (BU) strategy is more fine-grained since it does identify such categorical differences. Because the strategies were relativized to levels, these points must be interpreted against the background of a fixed level. A (BU) explanation at a higher level may of course be more coarse-grained than a (TD) at a lower one, but that does not invalidate the point that, at the same level, a (TD) explanation is more coarse grained than a (BU) explanation.

I will use possible worlds to characterize the difference between the two sorts of information that the two strategies provide. Imagine a set of possible worlds in which the same artifact (or its counterparts, for friends of counterpart theory) has slightly different categorical properties in each possible world. In the actual and all possible worlds, this artifact (and its components) will have a vast set of dispositions, $B$, containing the behaviors it would show in response to all sorts of external stimuli. In the actual world, a proper subset of this set, $B_f$, contains the dispositions that are relevant for the function it has (I am assuming the artifact functions well). Obviously, in some possible worlds the artifact will not have all the dispositions in $B_f$ because its categorical properties do not support them, but, equally obviously, in many other possible worlds the artifact does have all the dispositions in $B_f$ because the underlying variation in properties is irrelevant for that subset of behaviors. In those worlds, the artifact can still perform the function it has in the actual world, although it varies to

some extent from our actual artifact in its other dispositions and categorical properties.

With this model, the difference between the two strategies can be expressed more clearly. The (TD) strategy gives an explanation that is true in all those worlds in which the artifact can show the behaviors in $B_f$. To adopt Jackson and Pettit's (1990) phrasing, it picks out a trans-world pattern of relevantly similar worlds and shows how the actual world is but one of the possible instantiations of this pattern. It makes clear how the actual world fits in a pattern of possible worlds. By focusing on the similarities between the actual world and other possible worlds, the (TD) strategy shows us constancies between how things are and how they might have been. The (BU) strategy provides an explanation that is true in the actual world, but not in other relevantly similar possible worlds – in which the artifact also has the dispositions in $B_f$ but not the exact same underlying categorical spatial and physicochemical properties – because it provides all the details on the structural make-up of the artifact. It does not focus on similarities, but homes in on differences; it makes clear how the actual world differs from other possible worlds. It describes the particularities of the actual case and so differentiates the actual world from other possible worlds. Hence, the two strategies provide different sorts of information about the artifact. The (TD) strategy delivers *modally comparative* information about trans-world similarities; the (BU) strategy delivers *modally contrastive* information about trans-world dissimilarities (cf., again, Jackson and Pettit 1990, 1992a, b).

It is worth emphasizing that the relativization to levels of nature I talked about before has ramifications here as well. The relativity to levels of nature or organization of the distinction between the (TD) and (BU) strategies means that the information provided by them can come in various degrees of abstractness, which in turn means that the sets of possible worlds will be of varying sizes. A more abstract (categorical or dispositional) description of an artifact will be true at a greater number of possible worlds, since more abstractness entails greater tolerance of underlying variations. So in thinking about the sorts of information that the two strategies deliver, we have to keep the level of description fixed and apply the distinction between dispositional and categorical, as appropriate at that particular level. This procedure can be repeated at other levels.

### *6.4.2 Two Arguments for the Autonomy of Top-Down Explanations*

Perhaps someone feels that all I have said so far only adds up to her suspicion that (TD) explanations are really just deficient (BU) ones. After all, what reasons could there possibly be for regarding an explanation that contains the same type of information as another, only less of it, as self-contained? My arguments for the autonomy of (TD) explanations provide an answer to precisely this question. I will give two arguments that aim to establish the in-principle autonomy of

(TD) explanations. The first argument is in fact made up of two sub-arguments that depend on a specific account of explanation. The second argument also has two sub-arguments, which both show that (TD) explanations are better than (BU) explanations at meeting a very reasonable and intuitive requirement on what explanations must do. For someone who refuses to accept these arguments for in-principle autonomy, I will then show that there are definitely very strong pragmatic reasons for distinguishing my two strategies and granting them at least some amount of pragmatic autonomy.

First, general arguments in favor of higher-level explanation (such as those in the appendix to this chapter) apply equally well to the (TD) strategy. Let me review two of them briefly. From the perspective of a unification account of explanation, high-level explanation is often more unifying because it provides a unified treatment of phenomena that look different from a lower-level perspective but are relevantly similar at the higher-level. One and the same explanation applies to different low-level phenomena and that increases unification. The same holds for the (TD) strategy: It can provide more unification by explaining artifacts with different underlying categorical properties but (roughly) identical relevant dispositions in the same way. It applies to a broader range of artifacts because it omits underlying details. (BU) explanations, however, tell a different story for every variation in underlying categorical properties, exactly because they mention all these underlying details. As a result, they are less unifying.

Next, from the causal perspective, the question is whether the top-down strategy does a better job at picking out the explanatorily relevant causes. Here, too, an affirmative answer seems very reasonable. The explanatorily relevant causes of an artifact's behavior are more likely to be identified in terms of sub-behaviors than in terms of the underlying categorical properties, again precisely because of the tolerance (TD) explanations have for irrelevant underlying variations. As long as underlying properties realize the same behaviors, it does not matter if they differ in details. That, I think, is a good reason for believing that the behaviors, and not their underlying categorical properties, will come out as the explanatorily relevant causes in many relevant contexts.

So specific theories of explanation support the claim that (TD) explanations are perfectly good and self-contained explanations of an artifact's behavior, and not just wanting stories that ought to be supplemented with the extra information provided by the (BU) strategy. One drawback of this argument is of course its dependence on a specific account of explanation. Therefore the second argument tries to establish its conclusion independent of a particular account of explanation, starting just from seemingly reasonable and intuitive requirements on the task of explanations.

It seems that, no matter what your favorite theory of explanation says, it is a

desirable feature of explanations that they are robust and stable, i.e., that they apply across a fairly broad range of situations and are insensitive to irrelevant underlying variations. Explanations should cite the crucial factors and not every arbitrary detail about a situation. As Garfinkel puts it: 'When we explain an event, not everything about it [i.e., not every low-level detail, JdR] is essential to it. The explanation has to be stable in some neighborhood of the actual world.' (1981: 32). This requirement is better met by (TD) explanations than by (BU) ones. The former are more stable under perturbations than the latter, since they are true in a bigger region of possible worlds surrounding the actual world. With the above characterization of the sort of information provided by the two strategies it is easy to see why. (TD) explanations are insensitive to irrelevant underlying variation in categorical properties. They remain true regardless of possible variations in underlying structural details. As long as the underlying mechanisms and components make the same behaviors possible a (TD) explanation holds good.

Another intuitive requirement on explanations is that they make clear *what could have been different* (cf., e.g., Garfinkel 1981; Strevens 2004a). In a way, this is an extension of the stability requirement. The idea is that explanations should support the right kind of counterfactuals; the factors cited in the explanation should be those that make a difference to the explanandum's occurring. If a particular factor is cited as part of an explanation, that normally carries with it the suggestion that, had that factor not been present, things would have gone otherwise and the explanandum would not have occurred, at least not in the way in which it now occurred.

This desideratum follows analytically if one embraces a causal theory of explanation and a counterfactual (Lewis 1973) or INUS (Mackie 1974) account of causation, but it can also be motivated on independent grounds. One of the chief functions of explanation has often been thought to be to enhance our capabilities for prediction and control (classic sources for this point are Hempel and Oppenheim (Hempel and Oppenheim 1948) and Hempel (1965a) but Woodward (2003) also makes much of it). If the tradition is correct in forging this link, then the requirement that explanations support the right sort of counterfactuals stands to reason, for information about what could and could not have been different clearly increases our abilities for predicting and controlling events.

Applying this requirement to the two strategies again bolsters the autonomy of (TD) explanations. The critical factors that induce qualitative change are the behaviors, not the way in which they are realized. That is one of the morals of multiple realization. The underlying properties categorically described may be metaphysically abstruse beyond anyone's imagination, as long as they make the same behaviors possible it just does not matter. Because it only mentions

dispositions and their manifestations the (TD) strategy suggests the right kinds of counterfactuals: Had this behavior not been in place, the artifact would not have performed its overall behavior well or at all. The (BU) strategy, in contrast, incorrectly suggests that the spatial and physicochemical properties in their categorical descriptions make a crucial difference to the artifact's overall behavior. This is, of course, not to say that these underlying properties are totally irrelevant, for that is plainly false. Whether a bridge is made of concrete or green jelly is indubitably of crucial importance to its overall behavior. The point is that categorically described properties do not reliably pick out the dimensions along which instability and qualitative change occurs. Some changes in categorical properties lead to qualitative change in functioning of the artifact, while others do not. The (BU) strategy by itself does not tell you this, because it provides so much additional details. The (TD) strategy does carve out these dimensions of qualitative change. If behaviors change, the artifact will cease to be able to show its overall behavior (in the right way). Take a simple example: From the (BU) perspective the diameter of a screw can vary continuously without there being any qualitative difference between one or another diameter. Once we take the (TD) perspective into account, however, and learn that the behavior of this screw in a particular (type of) artifact is to fasten one component to another, by fitting in a screw hole of 6 mm., it will matter greatly whether this screw has a diameter of something very near 6 mm. – any other diameter will just not do. Another example: There is nothing special about the spatial properties and specific molecular structure that give rise to a particular elasticity coefficient of a spring as opposed to other spatial and microstructural properties, but if this spring is to exert a certain minimal force in a given artifact, there will be a boundary of qualitative change in the continuous range of possible elasticity coefficients and this boundary is precisely the one that is picked out by the behavioral description of a (TD) explanation.

To wrap up this argument, let me repeat the central claim succinctly. The (TD) strategy, and not the (BU) strategy, shows you what could have been different and that is an important explanatory virtue. The (BU) strategy cuts things up too fine-grained; it does not bring out the important fact that some classes of changes in underlying properties are irrelevant to the explanation because they do not affect the dispositions that do the explanatory work in understanding the artifact's overall behavior. For this reason, a (TD) explanation is self-contained and not incomplete.

### *6.4.3 An Additional Pragmatic Argument*

I think the above arguments for the conclusion that (TD) explanations are good, complete, self-contained explanations are compelling. I think they show

that we need to distinguish the two explanatory strategies for theoretical in-principle reasons. But perhaps not everyone is convinced. One may think that unification and causal accounts of explanation are on the wrong track so that any argument that depends on them is flawed. Or that the requirements I put forward as reasonable and intuitive are in fact wrongheaded. Or that my reasoning from them is flawed. In that case, I will argue in this section, there are still strong pragmatic reasons for differentiating between the two strategies and considering (TD) explanations as (pragmatically) self-contained. The arguments in this section thus aim to establish the claim that there are plenty of contexts in which only the (TD) strategy and not the (BU) strategy is called for.

First of all, if one insists on taking a hard-line reductionist position, all of the above considerations could be reinterpreted as pragmatic principles governing our practices of explanation. That is, if one really thinks that the only genuine explanations are provided by fundamental physics and that all the special sciences deal in temporary stand-in accounts, one could take all of the above as useful insights in how we treat these temporary stand-in accounts. On such a view, however, even the (BU) strategy will have to be replaced by a more detailed account about sub-atomic particles and their properties. Let us leave this extreme view for what it is, while duly noting the point that everything I have said can be reinterpreted as insights about the pragmatics of explanation.

A second pragmatic consideration is that the object of explanation of the (BU) strategy is too specific. Often enough, when someone wants to know how an artifact works, she is only interested in how the complex behavior exhibited by the artifact can be realized by simpler sub-behaviors and not necessarily in how these sub-behaviors in turn are made possible by the precise categorical properties of the artifact and its components. In such cases, the (BU) strategy is hyper-specific by providing information that has not been asked for. More specifically, we can think of use contexts and engineering design contexts. An interested user might want to know how the behavior that is the function of, say, a television set is realized by the multifarious behaviors of the set's components, but engineers designing a new television also need to know the categorical physical and chemical properties that explain how the components can show these behaviors. In fact, part of their task is to find just those materials and components that will show the behaviors they have in mind as efficiently and effectively as possible. That, it seems to me, is an excellent reason for them to be interested in all the details provided by a (BU) explanation.

The point about the different sorts of information provided by the two strategies can also be pursued in other directions. Earlier, I contrasted the two strategies by arguing that the (TD) strategy provides modally comparative information while the (BU) strategy delivers modally contrastive information. If that is correct, we have another reason to accept the pragmatic autonomy

of the (TD) strategy. It seems perfectly reasonable that someone who wants an explanation of how an artifact works is just interested how the working of the actual artifact is similar to that of other possible realizations of that artifact and not in how this one differs from other ways it might have been. The (BU) strategy swamps the modally comparative information in modally contrastive details.

Similarly for the counterfactual information suggested by each strategy. I argued above that the (TD) strategy does a better job at suggesting the right kinds of counterfactuals because it mentions factors that are crucial to the artifact's behavior, as opposed to factors that are merely of some influence. The (BU) strategy does not; it gives so much details that one loses sight of what could have been different. This is another pragmatic reason to think of the (TD) strategy as at least pragmatically self-contained.

So even if one thinks that (TD) explanations are not self-contained in principle, there is still plenty of reason to think they are so in a more pragmatic way. Contexts in which a full (BU) explanation would be explanatory overkill abound. Very often a (TD) explanation without the full categorical underpinning suffices.

### 6.5 Caveats and Objections

Before I continue and compare my proposal to other recent work on mechanistic explanation, I want to dismiss two possible objections.

First, even though someone may be inclined to admit that the (TD) and (BU) strategy are two self-contained strategies, she may nonetheless think that I am exaggerating the difference between the two. When we look at everyday artifacts and engineering practice, it will typically be the case that artifacts with the same functional decomposition will also have the same underlying categorical properties. A particular sub-behavior is virtually always realized in one particular way, with one specific sort of material device, because that realization just happens to work really well so that there is no need to look for different realizations. Hence, all this talk of different sorts of information, different supported counterfactuals, and different dimension of change is highly puffed up.

This is an interesting suggestion but it does not constitute an objection to my arguments. For two reasons. First, it is an open empirical question whether functional decomposition and physical realization go together so tightly. This may indeed be the case for some artifacts with complex components that have to satisfy a very specific set of behavioral constraints, but when we look at simpler artifacts – say, pens, cups, staplers, or switches – there is considerable variation in material realizations implementing the very same functional decomposition. Underlying categorical properties can differ substantially even when they all exhibit the same required behaviors.

Second, the objection is also slightly beside the point in that it seems to assume that the argument hinges on *actual* variation of underlying categorical properties. That is not the case, however. The arguments only require *possible* variation. Nothing in what I have said implies that there must be actual variation in underlying categorical properties; it suffices if there *could be* variation, because all the arguments rely on modal considerations. For example, the argument based on the stability requirement for explanations says that (TD) explanations are more stable than (BU) ones because they are true in a greater nearby region of possible worlds. Similarly, the argument that (TD) explanations do a better job at supporting the right kind of counterfactuals turns on the requirement that an explanation show what could have been different, as opposed to what is actually different in various instances. Hence, *possible* variation in underlying properties is sufficient for the arguments to go through and the objection that there is no actual variation fails.

All the worse, my imaginary opponent may retort, because that only proves that these stability and counterfactual support requirements are misguided. Such a radical rejection would be odd, however, because these requirements are really rather intuitive (cf. also the appendix to this chapter). The idea that more stable explanations are preferable over less stable ones makes perfectly good sense on a modal interpretation that takes into account how things might have been. Since, for all we know, how any high-level event actually happened is but one of many many ways in which it might have happened (i.e., might have been realized by lower-level happenings), it is a virtue for an explanation if it is insensitive to those possible underlying differences. An explanation that shows how to explain a given type of high-level event regardless of the exact way in which it is realized at the lower level, is desirable given our interest in high-level events. An analogous story can be told for the requirement that explanations have to clarify what could have been different. Since we are interested in prediction and control, an explanation that tells us what could and could not have been different is to be preferred over one that does not. Besides, knowing what were the crucial difference-making factors for an event's occurring also seems intrinsically valuable from an epistemic point of view as it gives us insight in how the world works – it cuts the world at its joints, so to speak. Here too, for all we know, the actual way in which any event happened was just one of many ways in which it could have happened. Therefore, an explanation that picks out the decisive difference makers for these types of events is superior to one that mentions all sorts of accidental features of how this particular event happened. So we see that both requirements rely intrinsically on modal considerations and that that is exactly why they make so much sense. It would be crazy to reject them just because one does not like the fact that some artifacts do not come in different varieties of underlying categorical properties.

The second worry is more of a caveat. It might seem that an explanation along the lines of the (TD) strategy is susceptible to something like the old dormitive virtues objection. (Why does this dormitive put one to sleep? Well, because of its dormitive virtues. Generalizing: why does *x* have effect *e*? Because it possesses *e*-producing powers. True enough, but not informative as an explanation.) An explanation of how a paperclip works that only refers to the paperclip's paper-clipping disposition surely cannot count as a serious explanation.

I agree, but that does not show that there is something wrong with the (TD) strategy; it only shows that it has its limits and that it must be applied in the proper manner. Specifically, explaining behavior (TD) style is explanatory when the behavior to be explained is complex, i.e., decomposable into sub-behaviors that are in some sense simpler and more basic. This also means that the objection is slightly misleading; a (TD) explanation does not explain an overall disposition to $\phi$ by the artifact's $\phi$-ing powers, but by sub-behaviors that are carried out by the artifact's components. And while the former is indeed rather uninformative, the latter is not. What the objection does bring out is that (TD) explanation stops at some point. When behaviors are basic, i.e., considered to be insusceptible to further decomposition, the (TD) strategy has nothing to contribute to our understanding of how that behavior is made possible. In that case, the (BU) strategy takes over and explains the basic disposition by a categorical description of underlying properties.[80] When one wants to know, say, why a sharp knife can cut, there is nothing to be learned by an explanans that only mentions the knife's cutting powers. In such cases, (TD) explanations are indeed impossible. The cutting disposition can be explained (BU)-style by the hardness of the metal and the spatial properties of the blade. What this shows is not that the (TD) strategy is of dubious stature, but that it only works for non-basic behaviors. That, however, is hardly a surprising conclusion, let alone a troubling one.

At this point, the primary project of this chapter is completed: the development of an account of artifact explanation. In the remainder of the chapter I will look at recent work in the tradition of mechanistic explanation, for there has been something of a revival of that tradition in recent years. I think my artifact-inspired account of mechanistic explanation has something to add to this general debate. What follows is tangential to the main project of this thesis. Readers who are exclusively interested in philosophy of technology may as well skip to the next chapter.

[80] Recall that basicality of behaviors and dispositions is relative to engineering fields and design tasks. This relativity carries over to the present point. A behavior that counts as basic in one field may be further decomposed in other fields.

## 6.6 Mechanistic Explanation

At the beginning of the chapter I indicated the connection of my work with the tradition of mechanistic explanation. In this section I want to explore this connection further by contrasting my work with that of recent authors in this tradition. It will turn out, unsurprisingly, that a large part of it is very congenial to what these authors have to say, but, more surprisingly, I think my construal also reveals a blind spot in their work. Before getting into that, I will present a brief overview of recent ideas about mechanistic explanation.

### *6.6.1 Overview*

A good place to start is Stuart Glennan's (latest) definition of a mechanism:

> A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations (Glennan 2002: s344).

For Glennan this definition captures a number of important aspects of mechanisms. First, that it only makes sense to talk about mechanisms as being *for* behaviors, not mechanisms *simpliciter*; mechanisms are intrinsically relativized to particular behaviors. Second, mechanisms consist of material parts or objects; they are not just sequences of processes but consist of things that do things. Third, mechanisms produce their behavior by the causal interaction of their parts. The interactions must be invariant in Woodward's sense (Woodward 2003), meaning roughly that they should support a relevant range of counterfactuals.

Glennan puts his definition to work in an account of mechanistic explanation. He distinguishes between the explanation of singular events and general regularities. Only the latter is relevant for my purposes, since explaining how an artifact works or is supposed to work is to explain a general regularity. It automatically is an explanation of how artifacts of this type in general work. To explain a regularity one needs a mechanical model: 'A description of a mechanism, including (i) a description of the mechanism's behavior; and (ii) a description of the mechanism which accounts for that behavior' (Glennan 2002: s347). This mechanical model in fact is the explanation; (i) is the explanandum and by virtue of showing why (i) is correctly ascribed to the mechanism in question, (ii) forms the explanans. A particular behavior that needs explanation is explained by a description of the mechanism for that behavior, which includes a description of how the parts of the mechanism, triggered by an outside event, interact sequentially to produce the behavior being explained.

In another recent paper Peter Machamer, Lindley Darden and Carl Craver (2000) defend what they claim is a markedly different account of mechanisms.[81] They go through some pains to argue that mechanisms really, i.e., ontologically speaking, consist of two different kinds of things: entities and activities.[82] However this may be, their definition of a mechanism reads: 'Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions' (Machamer *et al.* 2000: 3). Descriptions of mechanisms explain by rendering the explanandum intelligible. They do so by showing how the set-up conditions and intermediate entities and activities produce the outcome, which they call the termination conditions.

In another paper Carl Craver (2001) describes how the account of mechanisms just described gives rise to an account of multilevel mechanistic explanation. For behavior *b* of an item, we can give (1) a *contextual* description of that behavior, (2) an *isolated* description, and (3) a *constitutive* description. The three correspond to perspectives on an item's activity: (1) places the behavior in a larger system with its own behavior to which *b* contributes, (2) is a description of *b* in isolation, and (3) shows how *b* itself is brought about by sub-behaviors of parts of the item. By employing these perspectives, science uncovers hierarchical mechanisms and behavioral roles played by the components of these mechanisms. Craver claims that such role ascriptions to items are important scientific achievements since they allow us to assign items their proper place in the multi-layered picture of how our world works. The levels he talks about are a matter of perspective; they do not correspond to levels of nature. Hence, there is no correct and unique way to do describe any particular behavior.

### *6.6.2 Some Critical Remarks*

I now want to present two related criticisms on these accounts that derive directly from what I have said above about mechanistic explanation. My complaints do not so much concern what these authors do say, but rather what they do not say. I am very sympathetic to their efforts, but I think there is room for refinement and improvement. In particular, and perhaps unsurprisingly,

[81] But compare Tabery (2004), who argues that Glennan's and Machamer *et al.*'s approaches complement one another.

[82] I am not sure what the advantages of this emphatically ontological characterization are supposed to be. Other characterizations, such as Glennan's, are silent about the exact metaphysics of objects, processes, and mechanisms, but appear to be perfectly compatible with whatever metaphysic turns out to be correct. It seems to me a bad idea to conflate an account of mechanisms and mechanistic explanation with metaphysical claims about the merits of process and object ontologies. Neither project is likely to benefit from this.

it seems to me that they should give more attention to the ways in which mechanisms and their behaviors can be described.

First, they do not distinguish between a categorical and dispositional way of talking about objects and their properties. As a consequence, they also overlook the difference between the (TD) and (BU) strategies and the sorts of information and understanding they provide. I guess this is due to a failure to analyze in any detail different modes of describing mechanisms and their behavior. For instance, Glennan's characterization of a mechanical model does not stipulate whether the description of a mechanism is to be given purely in terms of behaviors or also in categorical terms. The two options lead to different sorts of information about a mechanism's operations, as I have argued above.

Craver does take a stand on the issue; he is talking about functional roles – dispositions and behaviors, in my terminology – although he does recognize the importance of the spatiotemporal organization of the behaviors described. As a result his analysis fails to take into account how categorical properties (besides spatiotemporal ones) often explain an item's behaviors. We saw in chapter 4 how categorical information about an underlying implementation always permeates an explanation in the form of a functional decomposition. First, because most functional decompositions implicitly contain spatiotemporal information about the behaviors they describe. Second, because the specification of the described behaviors may suggest categorical information about how they are realized. And third, because any functional decomposition reflects choices to realize behaviors in a particular way to the exclusion of the other ways it might have been realized. The way in which behaviors are decomposed into sub-behaviors reveals these decisions. Craver only notes the first of these three points. Furthermore, because he only talks about behavioral roles, he seems to exclude the possibility that behaviors are explained by underlying categorical properties instead of lower levels of sub-behaviors. He only allows for multiple levels of behaviors. By overlooking the role of categorical descriptions in mechanistic explanation, he also overlooks the distinction between the two different strategies of mechanistic explanation.

Machamer *et al.* present a subtler example of the same confusion. They talk about hierarchical descriptions of mechanisms *bottoming out* in lowest level mechanisms, where the expression 'lowest level mechanisms' refers to components that are accepted as relatively fundamental or taken to be unproblematic for a scientist or group (Machamer *et al.* 2000: 13). At first, this may sound like a recognition of the (TD) and (BU) strategies of explanation, but it turns out that the fundamental mechanisms should again be taken as activities, and not as entities with categorical properties responsible for the basic behaviors being explained. As before, we end up with a purely behavioral hierarchy again that does not leave any room for a (BU) underpinning by categorically described

properties. The different styles of mechanistic explanation go unnoticed.

Another oversight stemming from the same source is the difference between physical and functional decompositions, which none of the above authors recognizes. When they talk about decomposition, they seem to think that it is some sort of structural decomposition into physical parts, perhaps from a functional perspective. They assume that a functional decomposition automatically maps onto a structural one. In one sense, of course, this is true because one is free to carve up an object into physical parts according to functional or behavioral criteria. In another sense, however, such an assumption is plainly wrong, because structural decomposition from a functional perspective typically does not identify an object's non-overlapping proper physical parts, since parts often play a role in the execution of more than one behavior. As I argued in chapter 4, functional decomposition also has a more literal sense as decomposition of an item's behavior.

To be fair, however, I must add a disclaimer. I might have erred in interpreting Machamer, Darden, and Craver's intentions here (I am more confident about Glennan). Their central notions of entity and activity do not belong to standard philosophical discourse and the authors unfortunately do not bother to give them precise meanings. Hence, it is possible that the notion of 'activity' must be understood to include both dispositional and categorical ways of describing objects and their parts. Or that 'entity' is supposed to stand for a categorical description of an object. In that case, their account would turn out to be much more similar to mine. My only remaining quarrel would then be that they do not systematically treat the relevance of the distinction between dispositional and categorical talk for their accounts of mechanistic explanation.

A second conspicuous omission in the accounts of mechanistic explanation under discussion, following logically from the first, is the role and character of (BU) explanation of behavior. Although this point is clearly related to what I said above, it deserves some separate discussion. None of the authors I have been discussing devotes any attention to how the behaviors they take to be part of mechanisms are often explained by categorical facts about the mechanism, described at some appropriate level of abstraction. (This is all the more surprising once it is noted that their examples do contain examples of (BU)-style explanation.) Instead, they seem to think that all there is to learn about a mechanism is the spatial and temporal organization of the entities and activities in it:

> One understands a mechanism by discovering its component entities and activities, and by learning how their activities are spatially and temporally organized. ... Understanding how a mechanism works is just understanding how one activity leads to the next through the spatial

> layout of the components and through their participation in a stereotyped temporal pattern of activities from beginning to end (Craver 2001: 60–61).

Taken at face value (and subject to the same disclaimer as above), this strikes me as incomplete. It is clearly true that knowledge of the spatial and temporal organization of mechanism parts is important; my strategies also include such information. But even with this knowledge it makes perfect sense to ask how these spatially and temporally organized behaviors are made possible by the categorically described characteristics of the objects and materials that make up the mechanism. An answer to this question seems to me to be a legitimate part of a mechanistic explanation, which adds to our understanding of the mechanism in question.

Craver might want to reply that understanding a mechanism is an iterative process and that the behavior of a mechanism's parts must be understood, not by explicating the parts' categorical characteristics, but by another, lower-level, mechanistic explanation that explicates the parts' parts and their behaviors. That is indeed in line with what I have said about the possibility of creating sequences of (TD) and (BU) explanations, but it does not diminish the fact that categorical information has an important explanatory role to play in such sequences. (BU) explanations are part and parcel of physics, (bio)chemistry, materials science, and many engineering disciplines. Disqualifying them by stipulation flies in the face of a host of serious scientific work.

We find the same omission in Machamer *et al*. (2000). In their discussion of bottoming out, they distinguish four types of fundamental activities for molecular (neuro)biology: (1) geometrico-mechanical, (2) electro-mechanical, (3) energetic, and (4) electro-magnetic. These four types of activities are what the fundamental entities (macromolecules, smaller molecules, and ions) engage in (ibid.: 14). This is all right, but even if we know everything about the spatial and temporal configuration of a bunch of macromolecules and the activities they engage in, it still seems to be an eminently reasonable explanatory question to ask how it is possible that this particular set macromolecules engages in, say, some type of chemical bonding. Moreover, chemistry provides answers to questions like these by explicating the molecules' categorical properties (e.g., its microstructure).

Machamer *et al*. may object that they do not deny that such explanations are possible, but that they merely wish to argue that they are not part of mechanistic explanation at a particular level or in a particular discipline – specifically, the molecular biological level studied by biochemistry. Their argument for this view is that scientists tend to regard certain entities or activities as fundamental and unproblematic and hence do not endeavor to explain them further; they pass the buck to other scientific disciplines. I do not think this is a compelling

argument – at least not for the conclusion that categorical information has no role to play in mechanistic explanation. It is true that particular disciplines, both in science and engineering, tend to regard certain behaviors and components as basic, i.e., not in need of further explanation or decomposition. But we also saw that dispositional and categorical descriptions occur at many levels and not just above or below a particular threshold. So whether or not a given discipline considers some behaviors and components as basic is not directly relevant to the explanatory role of categorical information. At every level, there will be instances of behaviors being explained by categorical descriptions of underlying properties, and this will typically include more than mere spatiotemporal orderings of sub-behaviors. There is no reason to think that explanations of behavior in categorical terms automatically cross some boundary between different scientific subfields. I lack the biological literacy to present compelling examples of this in biochemistry, but engineering disciplines provide plenty of examples and nothing in them depends on unique characteristics of engineering as distinguished from other scientific fields. For instance, fluid mechanics explains the behavior of objects immersed in fluids by reference to categorical physical properties of the objects and fluids. Aerodynamics is all about how objects behave in airflows of various sorts and explanations of behaviors use categorical spatial and physical details of the objects and airflows. Explanations of this sort figure in mechanistic explanations of artifacts all the time. It seems to me that nothing prohibits the generalization of this observation to mechanistic explanations in other fields. Hence, I conclude that categorical, (BU) style explanation of the behavior of components in a mechanism should be part of a fully-fledged account of mechanistic explanation, as well as (TD) explanation purely in terms of behaviors or activities. I hope the arguments I have adduced for my own account have made it clear what the benefits of these additions would be.

## 6.7 Conclusion

I have outlined a general account of an important type of explanations in technology: artifact explanations that show how a technical artifact is able to exhibit specific behaviors, such as that associated with its function. Crudely put, explanations that answer 'How does this thing work?' questions.

My account draws heavily on resources provided by the causal-mechanical tradition in the philosophy of explanation, but adds to this tradition as well by suggesting a refinement having to do with the distinction between dispositional and categorical descriptions of artifacts and their properties. The core of my account consists of two explanatory strategies, which I dubbed the Top-Down strategy and Bottom-Up strategy. Let me repeat the brief description of these two strategies here.

(TD) Describe $b$'s sub-behaviors $b_1, b_2, \ldots, b_n$, their sub-behaviors, and so on. Continue until the level of basic sub-behaviors. Include the spatiotemporal configuration of the behaviors. (Put succinctly: Give the functional decomposition of $b$.) Point out ostentatively which structural components take care of the basic sub-behaviors.
(BU) Describe $a$'s physical components $c_1, c_2, \ldots, c_n$ sub-components, and so on. Continue until the level of basic physical components. To the extent that this is possible, give the categorical bases for all of the $c_i$'s relevant dispositions. Then, also give the functional decomposition to show how the lower-level basic behaviors just accounted for compose into $a$'s overall behavior $b$.

I have offered several arguments for why these two strategies each produce self-contained explanations. In particular I have argued that it is not the case that (TD) explanations are intrinsically incomplete because they do not give all the categorical details about an artifact and its physicochemical make-up. The information they provide is explanatory in its own right and does not need to be supplemented by the extra information contained in (BU) explanations.

In the final section I tried to show how the distinction between these two explanatory strategies, and that between dispositional and categorical talk about objects on which it is based, can be a useful contribution to recent discussions about mechanistic explanation that seems to have been overlooked so far, or at least not explicitly recognized as a useful refinement.[83]

## Appendix: The Legitimacy of High-Level Explanation

In section 6.4 I helped myself to the assumption that high-level explanation is legitimate, i.e., produces self-contained and good explanations. In this appendix, I offer three arguments to back up this claim. All the arguments are available in the literature; I have only collected them to refresh the reader's memory.

The first argument derives from the unification account of explanation discussed in the previous chapter. It is forcefully presented by Kitcher (Kitcher 1984) for the case of biology. On the unification account explanations should be unifying, which is to say they have to exemplify argument patterns that can be used over and over again to explain a wide range of different phenomena. (cf. Kitcher 1981, 1989). Given that requirement, we can ask whether it is possible

[83] I am indebted to Michael Strevens and Erik Weber for comments on previous versions of this chapter. Michael Strevens's seminar on high-level explanation at Stanford University during spring quarter 2004 led me to a lot of relevant literature for this chapter. An early version of this chapter was published as a paper in *Studies in History and Philosophy of Science* (De Ridder 2006).

that higher-level explanations satisfy it to a greater extent than lower-level ones. The answer, according to Kitcher, is affirmative. He considers biological explanations of hereditary and developmental phenomena and argues compellingly that (1) sometimes, explanation at the higher cell (cytological) level is to be preferred over the lower molecular level because it can be applied over a range of cases that look rather different at the molecular level but are relevantly similar at the cell level and (2) some lower-level phenomena must be explained at a higher level because the lower-level descriptions do not provide one with the resources needed to see the salient features that lead to the explanandum-phenomenon. It follows that higher-level explanation is not only tolerable as a harmless addition to lower-level explanation, but sometimes even indispensable for a proper understanding of a phenomenon. Higher-level explanations sometimes do a better job at best unifying our beliefs by being more general and applicable to a broad range of phenomena that look altogether different from the lower-level perspective. When they do so, they are to be preferred over lower-level accounts of the same phenomena.

The second argument turns on considerations provided by the causal account of explanation. On a viable causal account of explanation, one will recall from the previous chapter, an explanation ought to give the explanatorily relevant causes of a phenomenon (Hitchcock 1995). The question, then, is whether higher-level descriptions could be better at picking out the explanatorily relevant causes of a phenomenon. Everything naturally depends on how the notion of explanatory relevance is cashed out, but I think something can be said without making very restrictive or idiosyncratic assumptions about this.

Let us assume that, in some minimally committal sense of 'exist,' high-level phenomena exist.[84] We tend to group certain classes of phenomena under common headings such as molecular bonding, cell division, viral infection, a material's plastic deformation, pain, speaking competence, market share, revolution, etc. Even though these phenomena may be wildly heterogeneous from lower-level perspectives, they possess relevantly similar qualities at a higher level – typically qualities that are of interest to us in some or other context. Since we are often interested in such higher-level phenomena, we are interested in their explanation and that, according to the causal tradition, entails that we should be interested in their explanatorily relevant causes. Exactly because of the underlying heterogeneity it is highly plausible that the explanatory relevant causes of higher-level phenomena will also be higher-level phenomena. Think of the collapse of the Tacoma Narrows Bridge as a result of wind-induced vibrations. Whatever the exact causes of the bridge's failure

[84] Again, I am working from a worst-case scenario. If, as a causalist, you think the world contains real levels and higher-level causation, the defense of higher-level explanation will be much easier.

were – apparently experts still disagree (Petroski 1994, chapter 9) – they are most aptly described at a higher level: wind, vortices, resonance frequencies, oscillations, force, stress, etc. Surely there exists a long and intricate low-level description of the causes of the collapsing in terms of molecules and atoms, but it is hard to see how any reasonable construal of explanatory relevance would pick out this low-level description as identifying the explanatorily relevant causes of the high-level collapsing event. The point is that the detailed low-level description that was true of this particular collapse is but one of a vast range of possibilities that could have constituted the same high-level phenomenon. Because of that, it is extremely counterintuitive to think that the highly contingent lower-level phenomena are the explanatorily relevant causes of the collapsing.[85]

The point is easier to argue if we adopt a specific notion of explanatory relevance such as Strevens's (2004a) difference-making criterion, which prescribes that a causal explanation should cite only those factors that *made a difference* to a phenomenon's occurring, i.e., only those things that played an essential role in bringing about the explanandum count as explanatorily relevant. Armed with this criterion it is easy to appreciate why higher-level phenomena are often best explained by other higher-level phenomena. More often than not, the specific configuration of low-level phenomena is inessential for the occurring of a higher-level phenomenon. The exact distribution and speed of physical particles was not at all essential to the collapsing of the Tacoma Narrows Bridge. What was essential however, was that this low-level configuration realized the high-level event of, say, winds corresponding to the bridge's resonance frequency, which led to strong vibrations that caused a cable band to slip and the bridge to fail.

It follows that causalists, as well as unificationists, have strong reasons to endorse the legitimacy of high-level explanation. Definitely as an autonomous addition to low-level explanation and in some cases perhaps even as a superior substitute for low-level explanation.

My third argument does not depend on a specific account of explanation. No matter what your favorite theory of explanation is, *robustness* or *stability* is a virtue for an explanation. A good explanation should hold true in a range of similar cases and should be able to accommodate insignificant perturbations of actual circumstances without losing its validity. In general and other things being equal, explanations that can be used to explain a range of more or less similar events are to be preferred over ones that are only true for one particular

[85] I do not deny that there can be contexts in which we would be interested in the low-level details; I am just saying that these contexts are rarer than the ones in which a higher-level approach is appropriate.

event. As Garfinkel puts it: 'When we explain an event, not everything about it [i.e., not every low-level detail, JdR] is essential to it. The explanation has to be stable in some neighborhood of the actual world.' (Garfinkel 1981: 32). Such stability under perturbations is typically what high-level explanations do, and low-level explanations do not, provide. Let me give an example to clarify this point. Suppose the fuel tank of a car exploded because that car crashed into a tree at high speed. A low-level account of this explosion must mention the rises in molecule velocities (i.e., rise in temperature), the oxygen and fuel molecules that started the explosion, the chemical reaction leading to the breaking and creating of molecular bonds, the subsequent propagation of the chemical reaction, etc. Such an account is highly unstable under perturbations: Had the car's speed been slightly different, or the tank's location, or the exact initial configuration of molecules, or the fuel level in the tank, the low-level explanation would have looked different. It does not recognize the fact that there is a lot of redundancy in this case: Had this particular oxygen molecule not reacted with that other fuel molecule, others would have, with the same higher-level result. The higher-level account does recognize this redundancy and mentions only the high-level events that led to the explosion, such as the car's speed being above a certain threshold, the amount of fuel in the tank being greater than some minimum, the placement of the tank. It thereby black-boxes the inessential low-level realizations of these events. As a result, the higher-level explanation is far more stable; it holds in a wide range of similar circumstances.[86] Hence, if one accepts the stability desideratum for explanations one must also accept the legitimacy of high-level explanation, because often higher-level explanations are more stable than lower-level ones.

Finally, the same argument can also be put in terms of possible worlds. High-level explanations hold true in a neighborhood of the actual world, while low-level explanations are only true in the actual world, because they are so specific. In arguing for the legitimacy of both low and high-level explanation Frank Jackson and Philip Pettit (1990; 1992a; 1992b) explicate the difference between the two sort of explanation as a difference in information. Low-level explanation provides *modally contrastive* information; by homing in on the

[86] How wide a range and how similar the circumstances? There is no fully general answer to this, but I think the solution must be sought in the presuppositions of explanatory questions. Any explanatory request has certain implicit presuppositions about the kind of circumstances it deals with. In the above example, it could be a presupposition that we are talking about cars, as opposed to motorcycles or trucks, or even about a certain brand or type of car (a Ford Pinto, say), or that the tree (or other object) was sufficiently bulky, etc. An explanans is neither fully general, nor completely unique. The exact extent of the circumstances in which it holds will have to be determined on a case-by-case basis through ferreting out the exact presuppositions of the explanatory request.

particularities of the actual world it reveals how the actual world differs from nearby possible worlds. This information lets you see what is special about the actual world, in comparison to other nearby possible worlds. In contrast, high-level explanation provides *modally comparative* information; it focuses on trans-world patterns and similarities between the actual and nearby possible worlds. Such information tells you how the actual world runs on a pattern that spreads across a range of possible worlds. The conclusion is the same: high-level explanation has its own distinctive right.

# Intentionality and Normativity in Artifact Explanations

## 7.1 Introduction

In this chapter I investigate the remaining aspect of the connection between function and structure in the context of explanation; the contrast between a description of an artifact's function being an *intentional* description and that of its physicochemical structure being a *non-intentional* or *physical* description. This brings us back to the overarching theme of this thesis, which was the relation between the intentional and the physical perspective on artifacts, as exemplified in the relation between function and structure.

The account of artifact explanation I developed in the previous chapter did not say much about the intentional perspective on artifacts. The only thing close was my suggestion that we are typically interested in explanations of how an artifact is able to perform its technical function as opposed to one of its many other behaviors. Artifact explanations cover only two of the three aspects of the function-structure gap. They explain abstract descriptions of behavior by means of concrete categorical descriptions of an artifact's structure. Hence, they deal with the dispositional-categorical and abstract-concrete aspects, but not with the intentional-physical aspect. Artifact explanations, as I construed them, do not presuppose intentionality in any sense. This is not in itself a bad thing; there are no obvious reasons for why an account of artifact explanation would have to include intentional notions or presuppositions by necessity. The Dual Nature claim does not imply that anything that can be said about artifact must always involve both perspectives. At the same time, however, it would be odd if there were no decent way of relating an account of artifact explanation to the intentional perspective on artifacts, i.e., to human intentions, actions, artifact use plans, technical functions, etc. It is a good thing, then, that this is not the case. A plausible and compelling story can be told about how artifact explanations can defer all issues about intentionality to a function theory. That story is the story of this chapter.

The guiding idea of this story is that, in the context of explanation, the connection between function and structure splits into two parts. Artifact explanations take care of the dispositional-categorical and abstract-concrete aspects of the function-structure divide, and a function theory handles the intentional-physical contrast by relating an artifact's actual physicochemical

structure with the intentions of its designers and users. A consequence of this idea is that intentions, actions, and the normativity of technical function ascriptions deriving from them have no role to play in artifact explanations as such. When artifact explanations are combined with an adequate function theory, however, they can become explanations of an artifact's technical function, i.e., of an artifact's *intended* behavior. They then inherit the intentionality and normativity of technical functions and are drawn into the intentional realm, so to speak. They become explanations of how the artifact's designer or users *intended* it to be able to fulfill its function. With this combination of artifact explanations and a function theory, we have explicated the full connection between function and structure of artifacts in the context of explanation. And that means we will have answered the second main question of this thesis.

I develop this position in discussion with Peter Kroes, who has argued for an alternative construal of the explanatory connection between structure and function. In his proposal, which I present in section *7.3*, the connection does not bifurcate. I will argue in section *7.4* that, as a result, the proposal runs into irresolvable trouble. In the course of this discussion my own position will emerge. In section *7.5* I present my alternative proposal in greater detail and section *7.6* shows how this proposal steers clear of the trouble that dogged Kroes's idea while at the same time providing a plausible explication of the full connection between function and structure in the context of explanation.[87]

Before we get to all this there is another worry about the role of intentionality in artifact explanations that I want to address. Since artifacts are made and used by people, it might be thought that explanations of how they work will inevitably have human actions or intentions in their explanans. In section *7.2* I will show that this worry is correct, but only in a qualified sense that does no harm at all to the claim that artifact explanations are unintentional. The last section of the chapter, *7.7*, sums up its main points.

## 7.2 Intentionality and Action[88]

We start with the role of human action and intention in artifact explanation. The claim I will defend is that they have no essential role to play in artifact explanations, at least not *qua* their being intentional. This might sound a little crazy, for it is blatantly obvious that there are explanations of artifacts that do have human intentions and actions as essential parts of their explanans. Of

[87] This critical discussion of Kroes's work was previously published as a separate paper (De Ridder 2007). The paper-like formatting and material that has already been covered in this thesis have been removed, but the gist of the arguments remains the same.

[88] Parts of this section also appeared in De Ridder (2006).

course I am talking about explanations of an artifact's origin; explanations of *why it is there*, to borrow Wright's (1973) famous way of putting it. It can hardly be more evident that intentions play a part in explanations of why artifacts come into existence. Technical artifacts are, by definition, objects that have been purposefully made for a practical purpose and from that it follows directly that their origin must be explained with human intentions, goals, and actions. Mugs exist because people want convenient containers for their coffee and other hot drinks; and perhaps also because, in some distant past, people discovered how container-shaped objects were really useful for holding liquids. Actions and intentions play a major role in such explanations. Pretty much the same will hold for any artifact.[89] My claim that intentions and actions are not essential to artifact explanations must therefore be understood more narrowly. It applies only to artifact explanations as I have construed them in the previous chapter, i.e., as explanations that account for an artifact's ability to perform its technical function or some other behavior on the basis of its physicochemical structure.

To mount the argument we need the account of design of use and design that I discussed in chapter 2 (Houkes *et al*. 2002; Houkes and Vermaas 2006). On this account, artifacts are always embedded in a use plan. This use plan consists of a considered series of actions, undertaken to achieve a practical goal and at least one of the actions in it involves the manipulation of the artifact. By doing something with an artifact, a change in the obtaining state-of-affairs is brought about that moves the artifact user closer to her goal, a desired state-of-affairs. The important insight needed for my argument is that, although this intentional context of a use plan is always present for technical artifacts, it is not strictly required in the explanation of an artifact's ability to show certain behavior. For that, we need only the physical movements that realize the actions of the use plan; the fact that these movements realize an intentional action is not important for this explanatory purpose.

You will recall the paperclip example from the previous chapter (if not, never mind and just think of how you use a paperclip). This example is convenient to illustrate my point because it is so commonplace and yet contains all the elements required to make my case. Nothing essential to my argument would be lost were we to move to more complex artifacts. In the paperclip case, the action involved is the slight bending of the loops by pressing the longest loop against the papers and then sliding the clip on. Although humans typically perform this action, it is beyond question that the paperclip would still work were a robot or zombie to perform the same action. I take this to show that the

[89] Intentions may play a slightly less dominant role in traditional handcrafted artifacts like baskets, pottery, cloths, etc. (see Preston 1998a; Ingold 2000). But even if their genesis may be more like a process of organic growth, their maker will still have *intended* to make a basket, pot, cloth, etc.

intentionality of the actions performed with artifacts does not really matter for our understanding of how an artifact works. The only thing that matters for explanatory purposes is the physical movements going on, regardless of whether humans, zombies, or robots bring these movements about.

A distinction from action theory between an action qua intentional action and an action qua physical event can clarify this point. When I beckon someone to come over, my hand moves along a trajectory. This physical movement of my hand is the action qua physical event, or *intrinsic event* associated with my action of beckoning (Ruben 2003: 44). To the extent that human actions are involved in artifact use, the intrinsic events associated with these actions alone suffice to explain the artifact's operation. To understand the physical operation of the artifact, 'how it works', we need only see what sort of physical processes are going on. Although it is surely of practical value to know which of these processes are typically brought about by human actions, it does not matter for the artifact's behavior if they are brought about by a robot or a skilful monkey. The intrinsic event is doing the explanatory work in artifact explanations and not the intentional aspect of the action.

There is a bit more to say about how actions and artifact behavior go together in artifact explanations by bringing in my analysis of the distinction between dispositional and categorical descriptions of artifacts. Since artifact explanations are explanations of how an artifact can exhibit certain behaviors, they are, on my analysis, explanations of specific dispositions of an artifact. Since dispositions are by conceptual necessity connected with subjunctive conditionals, we might also say they are explanations of how it is that artifact $x$ would $\phi$ were it to be $\chi$-ed in context-determined ideal conditions (turn back to chapter 2 for details). Both (TD) and (BU) style explanations of this break the overall behavior down into various sub-behaviors, all of which are connected with similar subjunctive conditionals.

With this conditional analysis, we can see clearly what role the intrinsic events associated with human actions play in an artifact explanation. They provide the antecedent events on which manifestation of the artifact's overall behavior or its sub-behaviors is conditional. The intrinsic events associated with human actions stimulate the artifact in appropriate ways for it to manifest its behavior(s). I bend the loops of the paperclip out of position and it exerts a force in the opposite direction that holds a set of papers together. Somebody speaks into the mouthpiece of Edison's phonograph, the diaphragm starts to vibrate and the indenting point indents the recording material. By manipulating an artifact in appropriate circumstances you trigger one of its dispositions; the intrinsic event associated with your action is the $\chi$-ing, which makes the artifact (or one of its components) $\phi$, provided it is used in the right sort of ideal circumstances.

The ensuing picture of the role of actions and intentions is as follows. They typically figure in explanations of how an artifact is able to fulfill its technical function or other behavior because people use artifacts intentionally to achieve their goals. The intentional aspect of these actions, however, is explanatorily superfluous because the intrinsic events associated with them do all the work, as is shown by the fact that the explanation would be just as good when a non-intentional agent would bring about similar events.

### 7.3 Kroes on Technological Explanations

Before I present Peter Kroes's thoughts on it, let me reintroduce the issue about the supposed intentionality of artifact explanations again in somewhat different terms than before. First, take the claims that artifacts have technical functions and that technical functions presuppose intentionality. These were two of the starting observations for this thesis. Having a technical function carries with it a normative force; in some appropriate sense of 'ought,' an artifact ought to be able to fulfill its technical function. This is true regardless of whether it can in fact do so; it may be the case that the intended behavior that it is the artifact's function is not one of the things it can actually do. This is the phenomenon of malfunctioning. Now combine this with the intuitive idea behind my second main question: that there is an explanatory link between an artifact's function and its physicochemical structure. Artifacts are able to fulfill their function in virtue of their physicochemical constitution and designing engineers can explain how this is so. It then seems but a small step to the conclusion that explanations that account for an artifact's function on the basis of its structure must also presuppose intentionality. Their explanandum – an artifact's ability to perform its technical function – presupposes intentionality and is normative because the property of having a technical function does and therefore it may well seem that its explanans must also involve intentionality and normativity. If so, there would be something inherently special about these explanations. They would combine physical and intentional elements in a unique way, quite unlike normal physical explanations and equally unlike normal intentional and social explanations. Artifact explanations would constitute an interesting new genre of explanations.

I think there is something seriously wrong with this line of reasoning. To show what, I will analyze Peter Kroes's (1998; 2000) account of technological explanations, which more or less embodies it. Scrutiny of his arguments shows how they run together a couple of different points about artifact functions and explanations in an unhelpful way. When the confusions are cleared up, it will become clear how my artifact explanations and the intentional perspective on artifacts go together smoothly without there being inherently special about artifact explanations. In this section, I present Kroes's arguments in more

detail. The following section contains internal criticism of his arguments. In section 7.4.2 I will argue that there is a more fundamental confusion underlying his arguments and show how we can dispose of this confusion by dividing the connection between structure and function in two separate parts. The rest of the chapter then serves to flesh out this constructive proposal.

Before I proceed a word of caution about terminology. It will soon become clear that Kroes's *technological explanations* differ from my *artifact explanations*. To avoid confusion I will consistently use the former expression for his construal and continue to reserve the latter term for the type of explanations I discussed in the previous chapter.

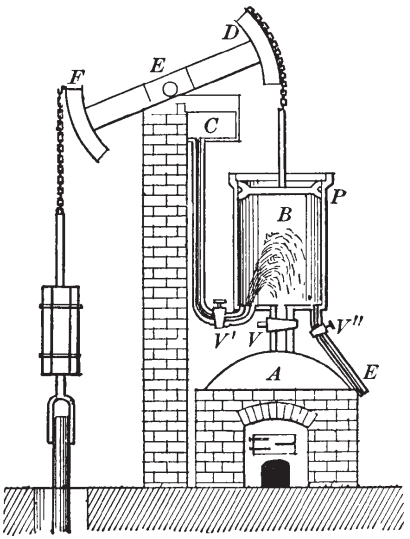
As Kroes construes them, technological explanations are explanations that explain an artifact's technical function on the basis of its physical structure.[90] Their explanandum is an artifact's having a technical function and the explanans is supposed to contain information about its physical structure. By means of these explanations Kroes aims to forge a direct link between an artifact's function and its structure, whereas I have been urging that this link has two parts.

To argue his point about the peculiarity of technological explanations, he first observes that technical artifacts have a dual nature. That should come as no surprise. We talked about this point at some length in the first chapter and there is no need to repeat what I said there. Artifacts can be described in terms of the function they have and in terms of their physical structure, with either description partially or wholly black-boxing the other. A clock is for time keeping, whatever its exact physicochemical make-up and, alternatively, someone without any experience with pencils cannot deduce that a 6-inch long hexagonal elongated piece of wood with a lead inside and one pointy end is for writing (though she might discover after a while that it can be used for writing). The two descriptions are logically independent: It is impossible to deduce an item's function from a purely physical description of its structure or its physical structure from just a description of its function. This means that standard deductive-nomological explanations are barred. D-N explanations are deductive arguments and there are no deductive inferences from a description of an artifact's physical structure to a description of its technical function.

To further flesh out his account of technological explanation Kroes presents a possible example of a technological explanation that involves the Newcomen steam engine. The main function of Newcomen steam engines was to drive water pumps. They did so by means of the up-and-down movements of their great beam. The great beam itself was driven by the actual steam engine that

[90] What follows here summarizes sections 4 and 5 of Kroes (1998).

***Figure 8*** *Newcomen's Steam Engine*




consisted of a boiler, a steam valve, and a cylinder with moving piston (see Figure 8). Roughly, an explanation of these engines has three parts.

(1) Physical laws or phenomena: e.g., steam occupies a larger volume than water, rapid condensation of steam in a closed vessel creates a partial vacuum, atmospheric pressure exerts a force on the piston.
(2) The physical make-up and configuration of the engine: the boiler, steam valve, movable piston, great beam, etc.
(3) Dynamic behaviors and causal interactions of the components: e.g., heating and expansion of water and steam, opening and closing of the steam valve, injection of cold water, condensation of water, creation of a partial vacuum, and movements of the piston.

Kroes correctly observes that it does not follow from an explanation along these lines that the technical function of a steam engine is to drive pumps, nor that it is to move the great beam up and down. All that follows is that the steam engine can be used to drive pumps, that it is a means to that end, or that it has the capacity to drive pumps. It is impossible to get the intentionality-presupposing normative explanandum containing the ascription of a technical function from the purely physical and descriptive explanans. The physical description of the artifact's structure does not fully explain the fact that the technical function

of steam engines is to drive pumps. Kroes concludes that the explanation as presented is not a real technological explanation since it does not properly account for the steam engine's function in terms of its structure.

Kroes (2000: 38–39) contains a sketch of a possible repair. Perhaps, he says, the relation between explanandum and explanans can be conceived in terms of pragmatic rules of action that are grounded in causal relations. For example, if one's goal is to drive a water pump, and a steam engine has the disposition to do so (i.e., something like the following causal conditional holds: Were the steam engine to be put to use properly in ideal circumstances, it would drive a water pump), then one can infer the following rule of action: To drive a water pump, use a Newcomen steam engine. In this context of action, the steam engine is a means to an end and acquires a function. The engine's physical structure still figures indirectly, since the rule of action is formulated on the basis of a causal conditional that was derived from the engine's structure. Kroes concludes: 'A technological explanation, therefore, is not a deductive explanation, but it connects structure and function on the basis of causal relations and pragmatic rules of action based on these causal relations.' (2000: 39).

In sum, Kroes's points are: (1) Technological explanations are explanations that explain an artifact's technical function by means of its physical structure; (2) because there are no deductive inferential connections between descriptions of function and of structure, D-N explanations cannot accomplish that task and that makes technological explanations special; (3) as it stands, the steam engine example is not a genuine technological explanation; and (4) by adding action rules to the example we do get a genuine technological explanation, so technological explanations connect structure and function by means of action rules.

## 7.4 Kroes's Arguments Reconsidered

In my opinion, both Kroes's arguments for the peculiarity of technological explanations and his account of them are flawed. In this section, I first offer two minor points of internal criticism. The first shows that the arguments do not establish that technological explanations are special, while the second shows that Kroes's repaired proposal is no more a satisfactory account of technological explanations than was his initial proposal. After that, I present a more fundamental critique, which aims to show that Kroes's way of setting up the problem is fundamentally mistaken.

### *7.4.1 Two Minor Points of Criticism*

I can be relatively brief about the first point. It is not quite clear which of the following claims Kroes is arguing for:

(1) A technological explanation of Newcomen's steam engine does not fit the mold of the D-N model of explanation.
(2) Technological explanations do not fit the mold of the D-N model.
(3) Technological explanations do not fit any of the currently available models of explanation.

To establish that there is something truly special about technological explanations, he should be interested in (3), because only that would support this conclusion. If the currently available accounts of explanation (such as the D-N account, unification accounts, and causal accounts) are capturing important aspects of what it is to be an explanation, and if technological explanations do not conform to any of these accounts, then they might represent an interesting new species of explanation worthy of philosophical attention. Unfortunately, however, the only claim Kroes establishes is (1). Of course, if (1) is correct and the modified explanation of the steam engine is a representative example of technological explanations in general, the truth of (1) makes (2) plausible as well. To the extent that this inductive argument is compelling, (2) would be supported as well.

But the plausibility of (2) does little to prove (3). For that, it would have to be shown that technological explanations fit none of the available accounts of explanation that we reviewed in chapter 5, e.g., Friedman's and Kitcher's unification accounts, Cartwright's *simulacrum* account, Salmon's, Lipton's, Strevens's, and other causal accounts, and Van Fraassen's pragmatic account. Even accounts of intentional explanation might be relevant if one concedes that artifact functions are intrinsically related to agents' intentions. Or accounts of social explanation if one is of the opinion that artifact functions are inherently social phenomena. I am not saying that this cannot be done, but Kroes has not done it. He has only shown that technological explanations cannot be construed as D-N explanations. While that may be perfectly true, it is hardly a reason for distress, since, for most philosophers of explanation, the D-N model has been relegated to the domain of philosophical relics.

In fact, as I will argue below, the situation is worse. There is every reason to think that his construal of technological explanations suffers from internal inconsistencies to such an extent that no account of explanation ought to fit it, on pain of being inconsistent itself.

My second point of criticism is that, as far as I can see, the suggestion to construe the relation between explanandum and explanans in terms of action rules is to no avail.[91] The step from a causal relation to a rule of action is relatively unproblematic: If one wants a certain effect and one knows one or

[91] In a recent paper Kroes acknowledges this point (2006).

more sufficient cause(s) of this effect, then, given the usual *ceteris paribus* clauses for causal relations and some hedging assumptions about the proportionality of the means in relation to the end, it is rational from a practical point of view to bring about this effect by bringing about one of these sufficient cause(s). If Newcomen's engine can drive a water pump if it is operated properly, one could use it to pump water if one wants so. But – and this illustrates the difficulty – in the same vein we can add that, if it can be used to tear stuff apart, one could use it to tear stuff apart if that is what one wants. One can use an electric guitar to play licks, and if one so desires, it would be rational to use it for that purpose, but if one is in a rock star type of mood, a guitar can also be used to smash loudspeakers, and it would be no less rational to use it to that end. None of this, however, goes to show that Newcomen's steam engine is for tearing stuff apart or that smashing loudspeakers is an electrical guitar's technical function.

Although the fact that something has a number of dispositions that are analytically connected to causal conditionals warrants inferences to various rules of action (under the assumptions mentioned), nothing singles out one of these rules in particular as somehow privileged and associated with the artifact's technical function. Neither does the artifact considered in isolation give you any reason to suppose that one of these dispositions is the artifact's technical function, as opposed to just something it can do. While causal knowledge may underpin rules of action, I do not see how it could sustain ascriptions of technical functions. At the end of the day the suggested repair is not much of an improvement over the initial proposal. All that can be inferred from action rules is that if a certain artifact can be used to accomplish some end, then it is rational to use it to that end, but that follows virtually analytically (again, given some background assumptions) from the fact that it is a *means* to that end, and that was already established in the initial proposal.

Hence, the project is unsuccessful so far. There are different possible diagnoses for this. Perhaps we shoud look longer and harder. Something else than action rules might be able to forge the required explanatory connection between structure and function. This is not my diagnosis; I think something is fundamentally wrong with how the project is set up. Unearthing this will make it clear that the project was a non-starter all along.

### *7.4.2 The Real Trouble and a Constructive Proposal*

I will start with a succinct statement of the fundamental confusion that vexes Kroes's project and then unpack it further by presenting an alternative proposal. The trouble is that Kroes stipulates that a technological explanation is an explanation that accounts for an artifact's technical function *in terms of its physicochemical make-up*. The way in which he tries to carry out his project suggest strongly that he takes this in the exclusive sense of *only* in terms of its

physicochemical make-up. In the papers we have been looking at he does not even mention – nowhere does he even mention a possible explanatory role for intentions, beliefs, social communities, etc. Given that functions are partially constituted by the intentions, beliefs, and actions of designers or users this construal of technological explanations is bound to fail. A technical function cannot be explained restlessly by just physicochemical structure, simply because it presupposes intentionality. Remarkably, Kroes endorses this point through his endorsement of the dual nature of technical artifacts, while not noting that it is inconsistent with his construal of technological explanations.[92]

I believe an alternative proposal immediately offers a correct diagnosis of what goes wrong in Kroes's project. It seems to me that his construal of technological explanation runs together two rather different projects, to wit (1) that of giving an account of technical functions and (2) that of explaining how, in virtue of its physicochemical structure, an artifact can fulfill the behavior that is its technical function. The result of (1) should be a set of necessary and jointly sufficient conditions for claims like 'artifact *x* has the function to $\phi$' being true or justified. It is fairly obvious that this set will contain more conditions than just those related to the *x*'s physicochemical make-up. In effect that is the negative result of Kroes's argument: Claims about technical functions cannot be deduced solely from information about the artifact's physicochemical structure. It is not quite so obvious that (knowledge of) a highly detailed account of *x*'s workings must be among these conditions, for that would mean that no one except highly knowledgeable engineers could ever be justified or correct in claiming that an artifact has a certain technical function. Project (2) should provide an account of how an artifact's physicochemical make-up enables it to exhibit the behaviors required for its functioning or other behaviors, as I argued extensively in the previous chapter. We saw there that a fitting construal of such explanations is to see them as mechanistic explanations. By decomposing the artifact's behavior they show how its overall behavior is made possible by a collection of sub-behaviors. In addition, they clarify how all these sub-behaviors can be carried out by the physical components of the artifact. What Kroes tries to do, however, is to get the results of both project (1) and (2) while drawing exclusively on the means for project (2). With just information about an artifact's structure, he tries to explain the artifact's technical function. That is an impossible task, because

[92] This has changed in the paper I referred to above. There he writes: 'More is needed than statements about structure, physical capacities, and practical reasoning for an underpinning of statements about proper functions of technical artefacts' (2006: 150). Unfortunately, though, he does not elaborate on this suggestion other than by suggesting that the missing element may be 'statements about social, institutional, facts.'

technical functions are inherently intentional and connected to a context of action, intentions, plans, etc.

An analogy will help to clarify the reason why. Suppose we want to explain why the function of the heart is to pump blood, or, more precisely, to determine whether it is true that the function of the heart is to pump blood or whether we are justified in asserting this. Surely, an elaborate scrutiny of hearts and their behavior by itself will not allow us to conclude that their function is to pump blood. Yet this is the only option open to us on an extrapolated version of Kroes's proposal, since he tries to determine what an item's function is just by looking at its physicochemical make-up. The antidote is to distinguish the project of spelling out the conditions for 'The function of the heart is to pump blood' being true or justified from that of explaining how the heart is able to pump blood. Accounting for the fact *that* the function of the heart is to pump blood is not the same as accounting for *how* it can pump blood. The first project will involve more than just the heart's intrinsic properties. Biological function theories disagree on exactly what more. Some suggest synchronic relational properties such as the heart's current contribution to organism fitness (Walsh 1996; Lewens 2004), while others look at diachronic relational (historical) properties such as the heart's ancestors contribution to ancestor fitness (Millikan 1984, 1993; Neander 1991a, b). The outcome of the second project, however, will look more like Kroes's proposed explanans. It will explicate how the physicochemical make-up of the heart and its constituent parts in their particular configuration leads to dynamic behaviors that, in the appropriate environment, add up to pumping blood.

The crucial lesson is that accounting for an item's function, on the one hand, cannot be done without taking the item's environment into account, be it its current ecological niche, its history, its ancestors, its users, its designers, or their intentions and beliefs. Functions are relational properties of an item, not intrinsic properties, and therefore they cannot be explained solely by looking at the item itself, isolated from its environment. An item's dispositions and its behaviors, on the other hand, are among its intrinsic properties and can be explained by looking just at the item's physicochemical constitution and mereological make-up.

Once we distinguishing these two projects carefully, we can see that Kroes's attempt to find a direct explanatory connection between an artifact's function and its structure was bound to fail. There simply is no such connection, because the real connection is between an artifact's function on the one hand, and its structure and its designer's or users' intentions, beliefs, and actions on the other.

## 7.5 A Constructive Proposal: Having vs. Explaining Functions

My alternative proposal thus is to divide the explanatory connection between an artifact's function and its physicochemical structure in two. A function theory explicates the conditions under which an intended behavior is the artifact's function, while an artifact explanation explains how the artifact is able, in virtue of its physicochemical structure, to show this behavior (for now assuming it is indeed so able). In this section I want to develop this proposal in more detail by looking at the relations and interactions between the two projects.

The first thing to note is that correctly ascribing functions and explaining how something can fulfill its function are largely independent. One can be justified or correct in claiming that so-and-so is an artifact's function without knowing how it is able to perform this function. Alternatively, one can explain how it is that artifacts (or their parts) have the dispositions they have or show the behaviors they show without knowing that one of these dispositions or behaviors is the item's function. Of course, one is typically interested in an explanation of how an artifact can perform the behavior associated with its function, since that tends to be its most interesting feature (that computers can function as paperweights is wonderful, but not usually why people care about them).

### *7.5.1 Relations Between the Two Projects*

The above should not be taken to imply that the projects are entirely unrelated; I have only argued that it is unwise to try to tackle them in one fell swoop. I now want to look at possible connections, two in particular. The first one is that an explanation of how something is able to perform the behavior that is its function might pop up in the justification for its having that function. Roughly, the intuition is that function ascriptions must have something to do with the actual dispositions of an artifact, or at least with our beliefs about the artifact's dispositions, where the beliefs in question have positive epistemic status (i.e., are justified, reliably produced, adequately grounded, or whatever else one thinks is required for positive epistemic status). In order to justify the claim '$x$ has technical function $\phi$' it is required that there be evidence that $x$ can in fact $\phi$. An adequate explanation of how $x$ can $\phi$ would be very good evidence, albeit not the only possible type of evidence. Naïve theories of artifact functions tend to overlook this intuition. Consider a theory that defines the function of an artifact to be what the designer intended the artifact to do. Such a theory lacks the evidence-requirement and thereby fails to link claims about functions to (evidence of) actual dispositions. As a result, it allows for crazy function ascriptions. A mad designer's intention to build a spacecraft from a bunch of matchsticks does not warrant the conclusion that the result he produces has the function of a spacecraft, for there is no way in which matchsticks could ever

compose a spacecraft, at least not by current scientific lights. So the first way in which the two projects are related is through justification. An explanation of how something can fulfill its function can be among the justificatory grounds for the claim that an artifact has that technical function.

The second connection is the one I talked about in the previous chapter and promised to discuss here. It appears most clearly in malfunction cases, but also holds for cases of well-functioning artifacts, as we will see below in section 7.5.2. For malfunction cases the second project I identified takes on a slightly different form, since the question of how the artifact can perform its function is obsolete when it cannot perform its function. What can be explained, however, and what is not obsolete, is how the artifact is or was *supposed* to perform its function. An answer to that question will look a lot like the answer to the original explanatory question, except that it will be phrased in terms of how the artifact's designers or users *intended* it to be able to perform its function. (The beliefs on which these intentions were based should have (had) positive epistemic status, or else the artifact would not really have (had) the function in question.) Such an explanation thus explicates how the designer intended the artifact's components to be configured, what dispositions she intended them to have, how she thought the components had these dispositions, and how she wanted the components to interact. Even if one does not think that this answer is valuable in and of itself, it is beyond doubt that it has instrumental value as background knowledge for determining the causes of malfunction. Only in contrast to how the artifact was supposed to work will it become possible to find out how exactly it malfunctions.

In this guise artifact explanations have a number of interesting features. First, they support counterfactuals about how the malfunctioning artifact would have worked, had the designer's intentions been successfully carried out and her beliefs been correct. In this way the entire explanation can be transformed into a big counterfactual claim: Had this (TD) or (BU) explanation been entirely correct of this artifact, it would have actually had the intended disposition that is its function. But more limited counterfactual claims are similarly supported. For instance: Had component $c_i$ had the disposition the designer believed it to have, the artifact would have been able to show overall behavior $b$; or: Had these two parts been in their intended places, they would have interacted so-and-so, thus yielding overall behavior $b$.

The second feature is that these explanations of how an artifact can fulfill its function also inherit the normativity of function ascriptions. This feature sets them apart from (mechanistic) explanations of natural phenomena. Although we might claim that photons 'ought' to behave as particles (as opposed to behaving as waves), this only goes so far as the theory from which we infer this claim has been inductively supported or as our previous experiences lend

inductive support to such a claim. Such claims express inductively supported expectations about phenomena. Artifact explanations, however, incur an extra and stronger type of normativity in that there are independently ascertainable objective facts of the matter as to how the artifact and its components ought to behave because there are independently ascertainable and objective facts about how the designer *intended* the artifact and its components to behave.[93] These facts are grounded in the fact that the artifact has a technical function for which it was intentionally designed. And this fact itself is grounded in the intentions, beliefs, and communicative actions of the designer(s) who devised the artifact (or in the intentions, beliefs, and actions of the community of users who put the artifact to a new use that has gradually become widespread standard use).[94] Under the assumption that she is competent and benign, i.e., is broadly rational and skillful, has beliefs with positive epistemic status about the materials and things she works with, and communicates veridically about what she intends the artifact's function to be, a designer objectively determines an artifact's technical function (see also chapter 2 for this). That fact entitles us to objective claims about what the artifact and its components ought to do, even in the face of malfunction.

Looking at the different kinds of grounds for claims about malfunctioning artifacts and 'misbehaving' natural objects can further elucidate the normative reading of artifact explanations. Designers base their beliefs about the artifacts they design on scientific theories or practical experience with the materials they use and in this sense their knowledge about artifacts parallels scientists' beliefs and knowledge about natural phenomena. The grounds for an engineer's claim that an iron bar ought not to buckle under a specific pressure does not differ in kind from those for the claim that a photon ought to behave as a particle; both are supported by normal scientific evidence. For any given artifact, however, anyone who did not design that artifact can adduce a different type of ground. Provided the competence assumption mentioned above is warranted – as it certainly is in our society – they can take the designer's word[95] as support for claims about technical functions and hence about how the artifact and its

[93] We may distinguish two types of cases: those in which we know in advance of the artifact's malfunctioning (because we brought it about ourselves or know that someone else did) and those in which we do not. In the former, we may still claim that the artifact or its components ought to do so-and-so, but this ought then does not carry an expectation with it that it actually can do so. In the latter type we do expect that artifact can do so-and-so and the artifact's malfunctioning truly surprises us because we did not know it before. While these are two slightly different senses of 'ought' I think they are both legitimate so there is no need to differentiate between the two further.

[94] For the sake of brevity, I will ignore such user-imparted functions for the rest of this section, but something very similar to what I am about to say can be said about them.

components ought to behave. This is a result of the epistemic division of labor in our society. Professional engineers and designers are entrusted with the task of designing correctly functioning artifacts for all manner of practical purposes. Non-designers have a social-epistemic right to expect designers to have beliefs with positive epistemic status or even mostly true beliefs about the workings and functions of the artifacts they make and to trust their testimony.[96] In other words, designers may be expected to be competent and benign. Failing these expectations leads to sanctions. This social-epistemic right has even been institutionalized most contemporary societies. Users have legal rights vis-à-vis producers of artifacts, which are laid down in product liability laws, warranty clauses and the like.

Because of all this, non-designers are entitled to objectively grounded normative claims about what artifacts ought to do. The grounds for such claims consist of beliefs with positive epistemic status or knowledge about what the designers intended the artifact to do. This knowledge or belief trumps other types of grounds, such as experience with the artifact, testimony about successful artifact use, or even a full-fledged artifact explanation of how the artifact is supposed to perform its function. Of course, non-designers sometimes also have these latter grounds for a claim that artifact $x$ ought to $\phi$, but my point is that they do not *need* it to be in a position to claim so. All they need for that is knowledge or belief with positive epistemic status that the designer intended $x$ to be for $\phi$-ing, i.e., that $\phi$-ing is its function. This screens off other types of justification and provides just the extra normative force that adheres to claims about technical functions. Precisely this normative force can be inherited by explanations of how a malfunctioning artifact ought to be able to fulfill its function.

To round off this lengthy explication of the normativity of technical function claims, let me give an illustration. Suppose I have been commuting happily in my car every day for the past year, but then one morning when I turn the key it does not start. I want to claim that my car still has its technical function (say, motorized personal transportation) and that it ought to start if I turn the key, even though it presently malfunctions. What sort of evidence do I have for this claim? Obviously my past experience with the car, but that is not crucial. What is more important is that I have every reason to believe that my car was designed and built by competent engineers with the purpose of designing and building something that has the function of providing motorized personal

[95] Or something derived from that through a chain of communication, e.g., what the label on the box says, or the salesperson, or your sister who just bought the same artifact.

[96] I owe this point to Wybo Houkes, cf. also Houkes and Vermaas (2006, especially chapter 8) and Houkes (2006).

transportation. Therefore, the normative claim that my car ought to start carries with it an extra and stronger normative force beyond that offered by the induction over past experience. Had I not had this experience, I would still have been entitled to claim that my car ought to start. Compare this with my successfully and regularly lulling my daughter to sleep by the monotonous humming sound of my car engine. If she will not go to sleep one day, I might also claim that my car ought to lull her to sleep, but this claim carries a different and weaker normative force because it lacks the additional support of claims related to technical functions. No designer of cars intended them to be for lulling babies to sleep and, as a consequence, I have no sort of special right to claim that they ought to do so.

### *7.5.2 Artifact Explanations of Technical Functions Reconsidered*

I want to develop the point about the intentionality and normativity of technical functions in one final direction. In the previous subsection I have been talking about how the artifact explanations of *malfunctioning* artifacts inherit the intentionality and normativity of technical functions. In fact, however, this inheriting mechanism is not limited to malfunction cases. Artifact explanations of well-functioning artifacts can also inherit the intentionality and normativity of technical function ascriptions.

To see why, consider what happens when a (competent and benign) designer gives an artifact explanation of a well-functioning artifact of hers. She then explains how this artifact is able to fulfill the behavior that is its function. Since the artifact can fulfill its function, the intended disposition that is its function is also one of its actual dispositions. Nonetheless, the explanation recounts how the designer wants the artifact to be able to do show this intended behavior and how she believes it can do that. Because her intentions have been successfully carried out and her beliefs about the artifact and its working are correct, this is also how the artifact works in actual fact. The designer's beliefs about the artifact and its properties form a transparent window on the real artifact, so to speak. The intentional version of the artifact explanation coincides completely with the non-intentional version that explains how the artifact has the disposition that happens to be its function.

Just as the intentionality of having a technical function is inherited, so is the normativity. The mechanism is again similar to that I discussed above for malfunction cases. Because of the social-epistemic role of the designer, her explanation is normative; it lays down how the artifact ought to work. Non-designers may point to this explanation as the objective truth about how the artifact ought to be able to carry out its function. Because the artifact functions well, this will also be how it actually is able to carry out its function.

At least in the vast majority of cases. Situations in which an artifact works

as intended but a designer is mistaken about how it does so are rare. The only artifacts about which designers might be mistaken will be highly complex artifacts or artifacts that rely on innovative but still poorly understood science. Especially for them, however, it is extremely unlikely that a designer would be successful at creating a well-functioning exemplar without having correct beliefs about its working. Nonetheless, there is a small possibility of the designer's explanation being incorrect during some period of time until the right explanation is discovered. In chapter 1, I mentioned the example of the steam engine. Early creators of steam engines apparently had a flawed understanding of the science behind steam engines and as a result their explanations of the steam engine's operation were partially incorrect. In these rare cases, the explanation still inherits the intentionality presupposed in the function ascription – after all, it describes how the designer intended the artifact to have the disposition that is its function. That such explanation would also support normative claims to the effect that this is how the artifact ought to be able to perform its function strikes me as counterintuitive. In what sense could it be correct to say that an artifact *ought* to have a disposition in a way that is physically impossible for it, even though we do not (yet) know this to be so? In these cases, the inference to normative ought-claims is therefore barred even though the intentionality inheriting mechanism remains intact.

The upshot is that artifact explanations of the intended behavior that is an artifact's technical function always inherit the intentionality and normativity that clings to the property of having a technical function. If an artifact functions well, its intended behavior is one of the things it can actually do. The explanation of how its designer intended and believed it to have this disposition then coincides completely with that of how it actually has this particular disposition. In addition, the explanation allows for a normative reading: This is how the artifact ought to have the disposition that is its function. If an artifact malfunctions, it does not have the disposition for which it was intended. In that case, the explanation of how the designer intended and believed the artifact to have this disposition is not completely true of the actual artifact. At least one of the behavior and property ascriptions in the explanation is wrong. The intentional version of the explanation does still support normative claims. It is still true that this is how the artifact ought to have (had) this disposition.

At this point the obvious follow-up question is whether the two projects I have been outlining in general terms are feasible in practice. It is one thing to formulate a set of requirements that a theory of functions and an account of explanation ought to satisfy, but quite another to show that these requirements can be met by an actual function theory and an actual account of artifact explanation. That is why I will use the next section to present a theory of

artifact functions that can be combined with my own account of artifact explanation to satisfy the requirements I have submitted.

## 7.6 Delivering the Goods

In chapter 2 I already put forward my basic assumptions about what technical functions are and what is required for an artifact to have a technical function. Technical functions, I said there, are the intended dispositions or behaviors for which an artifact is designed. Since artifacts are always embedded in use plans, we may also say that functions are the intended role of an artifact in a socially accepted use plan that issues from either the artifact's designer or a later community of users. Mere intentions do not suffice for bestowing artifacts with functions. To bar unwarranted and crazy function ascriptions, the beliefs about the artifact, its components, its materials, and their properties on which the intentions are based must have positive epistemic status. In addition, the designer's intentions must have been carried out reasonably successfully so that the actual artifact resembles what she had in mind to a large extent. In short, the designer or the users must have been *competent*.

### *7.6.1 A Theory of Technical Functions*

While these assumptions paint a fairly clear picture of technical functions, they do not amount to a precise analysis of the necessary and sufficient conditions for claims like 'artifact $x$ has the function to $\phi$' to be true. Since I have been urging in the previous sections that a function theory takes care of an important part of the connection between function and structure in the context of explanation, I think I owe the reader a somewhat more developed function theory.

Rather than developing my own theory, I will borrow the ICE-theory of technical functions developed by my colleagues Wybo Houkes and Pieter Vermaas (Houkes and Vermaas 2004; Vermaas and Houkes 2006; Houkes and Vermaas 2006). While this theory is a theory about correct or justified function ascriptions, rather than about the having of a technical function, it will suffice for the present purposes. I introduced the ICE-theory in the chapter 2 with the proviso that it needs some fine-tuning to handle malfunction better. We will get to that in a moment, but let me start with repeating its official statement. (For my remarks about it, refer back to chapter 2.)

> An agent $a$ [is justified or correct in ascribing, JdR] the capacity to $\phi$ as a function to an artefact $x$, relative to a use plan $p$ for $x$ and relative to an account $A$, iff:
> (I) The agent $a$ has the capacity belief that $x$ has the capacity to $\phi$, when manipulated in the execution of $p$, and the agent $a$ has the contribution

> belief that if this execution of *p* leads successfully to its goals, this success is due, in part, to *x*'s capacity to $\phi$;
> (C) The agent a can justify these two beliefs on the basis of *A*; and
> (E) The agents *d* who developed *p* have intentionally selected *x* for the capacity to $\phi$ and have intentionally communicated *p* to other agents *u*.
> (Vermaas and Houkes 2006: 9)

An important touchstone for any function theory is whether it handles malfunction cases well, i.e., whether the function ascriptions it warrants about artifacts that cannot perform their technical function are (by and large) in line with our intuitive function ascriptions to such artifacts. I believe the ICE-theory goes a long way, but if I am not mistaken condition (I) needs some fine-tuning if the theory is to deal adequately with all cases of malfunction. This does not affect the theory's general approach however. For the sake of brevity I will not go over a host of examples that the theory covers successfully, but limit myself to an outline of its strategy for coping with function ascriptions to malfunctioning artifacts and a discussion of one worry that leads to some modifications of condition (I).

That basic strategy runs as follows. Since agents only need justified beliefs, as opposed to knowledge, about the artifact's dispositions, the theory allows for cases in which an agent's beliefs are proven wrong by later evidence. This disparity between our beliefs and the reality they are about opens up a way to ascribe functions to malfunctioning artifacts. I may have every reason for believing my phone to have the right dispositions to allow me to call my mother and I may also satisfy all other conditions laid down by the theory and, by that token, be justified in ascribing the function of teleconversation to it, but if – unbeknownst to me – a practical joker has removed the microphone from my headset, it nonetheless malfunctions.

This example does raise a concern, for the theory seems to imply that once I have discovered that my phone malfunctions, I can no longer ascribe the same function to it because I no longer have the capacity belief that it has the capacity to transmit my voice to the other end of the line. That is a counterintuitive result. To deal with cases like these we must modify condition (I).[97] I propose the following: The agent does not have to have the belief that *x* can $\phi$ but may also have the overriding belief that the (competent and benign) designer intended *x* to be able to $\phi$ and that therefore *x ought to be able to* $\phi$ or that it *would have been able to* $\phi$, had particular interferences not occurred. The modified condition reads:

[97] The modifications I am about to suggest are in line with what Houkes and Vermaas say but not explicitly theirs.

(I') The agent *a* has the capacity belief that *x* has the capacity to $\phi$ when manipulated in the execution of *p*; or the belief that the (competent and benign) designer intended *x* to have the capacity to $\phi$ when manipulated in the execution of *p*, etc.

Of course, this belief of *a* must have positive epistemic status as well. With this modified condition in place, I can still be justified or correct in ascribing the function of teleconversation to my phone after discovering the removed microphone, for I can have a belief (with positive epistemic status) that its competent and benign designers intended it to be able to allow for teleconversation so that it would have had the right dispositions, had someone not been playing this joke on me.

Arguably, artifacts that no longer fulfill their function due to normal wear and tear form another class of malfunction cases. Even if one is hesitant to describe such cases in terms of malfunction – malfunction due to normal wear and tear may be considered part of correct functioning since no one expects artifact to last forever – it nonetheless seems that they are cases in which artifacts still have their function but can no longer perform it. A blunt knife still is for cutting, just as an empty pen is for writing and a worn out light bulb is for lighting a room. To also render justice to these and similar cases (I') has to be modified further. If an agent knows or has a belief with positive epistemic status that the artifact used to be able to $\phi$, but can no longer $\phi$ due to normal wear and tear, she is still correct or justified in ascribing the function to $\phi$ to it. This yields:

(I'') The agent *a* has the capacity belief that *x* has the capacity to $\phi$ when manipulated in the execution of *p*; or the belief that the (competent and benign) designer intended *x* to have the capacity to $\phi$ when manipulated in the execution of *p*; or the belief that *x* used to have the capacity to $\phi$ when manipulated in the execution of *p* but no longer has this capacity due to normal wear and tear, etc.

A final type of cases that might be thought to pose a problem for the modified ICE-theory are severely damaged or malformed artifacts. It seems counterintuitive that one could still be correct or justified in ascribing a function to a heap of scrap that used to be, say, a blow-dryer. Note that this is also a problem for the original ICE-theory. It is possible to have a false belief with positive epistemic status about what an artifact can do, while, unbeknownst to you, it has just been smashed to a thousand pieces. My modifications only open the door for more similar cases by adding two extra possibilities for correctly ascribing functions. This does not alter the nature of the problem, however.

I think the solution lies in denying that a severely damaged or malformed artifact still is an artifact, or at least that it still is the *same* artifact. While I am not able to give precise identity and persistence conditions for artifacts, it seems eminently reasonably to suppose that severe damage or malformation will cause the artifact to cease to exist. A heap of scrap that used to be a blow-dryer is no longer a blow-dryer and therefore cannot have the function of a blow-dryer. Although we may believe the blow-dryer still exists and satisfies all the modified ICE-conditions, our function ascription fails because the object to which we ascribe the function in fact no longer exists.

As far as I can see, then, the modified ICE-theory with condition (I'') in place does a satisfactory job of accounting for artifact functions. Since my project in this thesis is not to develop a definitive account of technical functions, we can leave it at this.

### *7.6.2 Combining the Modified ICE-Theory and Artifact Explanations*

The next task is to see if this function theory matches up with my account of artifact explanation in the way I envisaged. Not unexpectedly, I think it does. As I argued above, once a function theory has established that the intended disposition to $\phi$ is an artifact $x$'s function, an artifact explanation can explain how the $x$'s designer intended it to have that disposition. If $x$ functions well (and the designer is correct) this explanation details $x$'s actual physicochemical structure correctly. If $x$ malfunctions, the explanation is not completely true of the actual $x$, but correct in that it describes how the designer intended $x$ to have the disposition to $\phi$. That is exactly what my artifact explanations can do, as we saw in the previous chapter.

The crucial question is whether a combination of the ICE-theory of artifact functions and my account of artifact explanation works out. It should be obvious that there are no problematic issues if an artifact functions as it is supposed to. In that case, a (correct) function ascription according to the modified ICE-theory implies that the capacity and contribution beliefs are true and therefore that the artifact indeed has the disposition to $\phi$, which contributes to successful execution of the artifact's use plan. An artifact explanation then shows, (TD) or (BU) style, how the designer intended it to have this disposition. Since the intended disposition is also one of its actual dispositions, the designer's intentions and beliefs form a transparent window on the actual artifact and the intentionality and belief context can simply be stripped away to yield an explanation of how the artifact actually has this disposition to $\phi$.

If there are problems at all, they should arise for malfunction cases. When we ascribe a technical function to a malfunctioning artifact on the basis of the modified ICE-conditions there are three options: (1) we have a false belief that the artifact has the right disposition(s) to function correctly, (2) we have a belief

that a (competent and benign) designer intended $x$ to have the disposition to $\phi$ when manipulated in the execution of a use plan, or (3) we have a belief that $x$ used to have the capacity to $\phi$ when manipulated in the execution of a use plan but no longer has this capacity due to normal wear and tear. (In each case, our belief must have positive epistemic status.) In all three cases, it is still perfectly possible and useful to give an intentional version of an artifact explanation, which describes how the artifact was supposed to work, or how it was intended and believed to have the disposition to $\phi$ by its designers or later users who gave it a new function. Such explanations support normative claims about how the artifact ought to work. Or counterfactual claims about how it would have worked, had so-and-so interferences not occurred or had it not been worn out. And, for artifacts that no longer functions due to wear and tear, claims about how the artifact used to work. All of this works exactly as I said it would in section 7.5.1. Here are a few examples of what this could sound like to make the idea more vivid: This component should sit here and interact with that other one right there so that they would have shown such-and-so behavior, thus contributing to the proper functioning of the artifact. Or: this light ought to go on when I hit that switch. Or: this spring used to push that thing back.

I do not see any problems about intentionality and normativity here that an account of artifact explanation could not pass on to a function theory but would have to take on itself. By inheriting the intentionality and normativity that attaches to the property of having a technical function, artifact explanations can rest content in explaining how an artifact is able to fulfill its technical function.

Someone may worry about circularity in the combination of the ICE-theory and my account of artifact explanation, though. If part of the grounds for ascribing a function can be an artifact explanation (account *A* could be an artifact explanation) and if an account of artifact explanation lets the function theory take care of intentionality and normativity, doesn't that land us in some sort of grounding circle? Not if we look closer at the exact grounding relations involved. Professional engineers or technically savvy non-professionals will have at their disposal elaborate artifact explanations as grounds for the functions they ascribe to the artifacts they design. That means they will have mostly or only true beliefs with positive epistemic status about the dispositions and categorical properties of the artifact and its components (more accurately: the properties dispositionally or categorically described), the configuration and interactions of the components, and their dispositions and categorical properties. But the grounds for these beliefs, and hence for the explanation, in no way depend on the function ascription. Instead they are based on the normal grounding mechanisms for beliefs about the world: observation, experiments, experience, and testimony. So if a designer ascribes a function to

a malfunctioning artifact she has designed, the normativity of this ascription is in the end epistemic, derivative of the normativity of positive epistemic status. Although the grounds for a function ascription will rely on an artifact explanation for designers, the grounds for this explanation in their turn do not rely on the function ascription and therefore there is no circularity in grounding here.

Anyone who ascribes a function to an artifact she has not designed can acquire a belief with positive epistemic status by the normal mechanisms of belief and knowledge acquisition: observation, experience, testimony, or an artifact explanation (having seen, experienced, or heard that the artifact is for so-and-so, or knowing an explanation of how it can do so-and-so). In addition to the normative force of positive epistemic status, these laypersons have an additional objective normative claim that entitles them to say that an artifact *ought* to be able to fulfill its technical function properly, as elaborated in section 7.5.1 and 7.5.2. If you know or believe with positive epistemic status that an artifact's designer intended it to be for $\phi$-ing, then in virtue of the social division of epistemic labor between (professional) designers and non-designers you may claim that the artifact ought to be able to $\phi$. Whatever the details of these arrangements, we certainly do not stumble upon a grounding circle here and that is the point I wanted to argue.

The circularity worry is misplaced and the combination of the ICE-theory of function ascriptions and my account of artifact explanation can bear the burden. For all I can see, the two deal adequately with the normativity of technical function ascriptions. The ICE-theory pinpoints where the intentionality and normativity come from and gives the conditions for when function ascriptions to functioning and malfunctioning artifacts are justified or correct. All that is then left to do for my account of artifact explanations is to provide a general model for explanations of how an artifact can fulfill the intended behavior that is its technical function. These explanations inherit the intentionality and normativity of function ascriptions.

### *7.6.3 Connecting Function and Structure the Right Way*

In closing this section, I want to return briefly to Peter Kroes's original project of developing an account of technological explanations that explain an artifact's function by means of its physical structure. I argued that this project was doomed to fail in its original construal because it is impossible to explain an artifact's function *exclusively* on the basis of its physicochemical structure.

In the previous pages, however, I have implicitly collected the building blocks to create something that looks rather like Kroes's technological explanations, except that it involves more than just the artifact's function and its

physicochemical structure. In fact, I have even been describing that very thing in the present section. It is, of course, the combination of a function theory such as the modified ICE-theory and my account of artifact explanations. Together, these two explicate the full connection between an artifact's function and its physicochemical structure. In contrast with Kroes's construal, this connection is indirect; it runs through the intentions and beliefs of an artifact's designer or users.

A function theory identifies the disposition for which the artifact was intended. The artifact's actual physicochemical structure is brought into the picture because the beliefs on which the designer's or users' intentions are based are about the artifact's physicochemical structure, its components, their configuration and dynamic interactions. Because it is required that the designer or users be competent and benign, it follows that they will have mostly true beliefs with positive epistemic status about the artifact, which is to say their beliefs correspond with the artifact's actual physicochemical structure.

When we know what the artifact's function is, artifact explanations can explain how the artifact is supposed to be able to fulfill this function, i.e., explain how its designer intended it to have the disposition that is its function. They do so by explaining the abstract dispositional description of the artifact's intended behavior with a concrete categorical description of its physicochemical structure. Specifically, they either give a (TD) explanans that gives the functional decomposition of the artifact's intended behavior and allocates the basic sub-behaviors to its physical components, or a (BU) explanans that also provides the categorical underpinning of all basic behaviors. If the artifact functions well, this intentional version of the explanation also gives a correct explanation of how it actually is able to fulfill its function. If the artifact malfunctions, the intentional version of the explanation contains at least one discrepancy with the artifact's actual physicochemical structure and dispositions.

I am somewhat hesitant to call this conjunction of function theory plus artifact explanation a 'technological explanation' since it is definitely not an explanation in the sense of any of the available theories of explanation. Such terminological niceties should not keep us from noting that we now not only have a satisfactory solution to Kroes's problem, but also a direct answer to the second main question of this thesis. The above shows how artifact function and structure are connected in the context of explanation.

## *7.7* Conclusion

The general lesson to be learned from this chapter is that the connection between an artifact's function and its structure in the context of explanation divides neatly into two parts. One part, artifact explanations, takes care of

the dispositional-categorical and abstract-concrete aspects of the function-structure gap, while the other part, a function theory, handles the intentional-physical contrast of the gap.

To arrive at this conclusion, I first dismissed the worry that intentional action plays an indispensable role in artifact explanations. Actions will typically show up in an explanation of how an artifact can show the behavior that is its function, but the intentionality of these actions is explanatorily irrelevant. Only the physical movements, or the *intrinsic events* associated with the actions, matter when we are explaining how something works.

I then looked at how artifact explanations can be combined with a function theory to explicate the full connection between function and structure. To develop my own position, I discussed Peter Kroes's construal of technological explanations. His explanations unsuccessfully tried to forge a direct link between an artifact's function and its structure. I showed how that attempt was doomed to fail because Kroes refused to take the intentions, beliefs, and actions of the artifact's designer and users into account and tried to explain an artifact's function *exclusively* on the basis of its physicochemical structure.

My own alternative proposal has the following elements. First, the connection that Kroes tried to find between an artifact's function and its structure really has two parts and is in fact a connection between, on the one hand, an artifact's function and, on the other, its physicochemical structure *and* the intentions, beliefs, and actions of its designers, users, or both. One part is a function theory that explicates what it takes for an artifact to have a particular intended behavior as its function, and the other part an account of artifact explanations that shows how an artifact's dispositions can be explained by its physical structure.

Second, a function theory should account for the intentionality presupposed in the property of having a technical function, i.e., it should explicate how that property presupposes intentions, beliefs, and actions of an artifact's designer or its users.

Third, if artifact explanations are used to explain how an artifact has the intended disposition that is its technical function, they inherit the intentionality and normativity of that technical function. That is to say, they become couched in terms of what the designer *intended* the artifact and its components to do. If these intentions were successfully carried out and the beliefs on which they were based true, then this intentional version of the explanation coincides fully with a non-intentional explanation in terms of the artifact's actual dispositionally and categorically described properties. The designer's intentions and beliefs then form a transparent window on the artifact's actual physicochemical make-up and behavior. If something went wrong with the designers intentions, if her beliefs were partly false or

incomplete, if the artifact wore out, or if it broke down for some reason, then the intentional version of the artifact explanation is not completely correct of the actual artifact. At least one the dispositional or categorical descriptions of the artifact or its component must be false; there is a discrepancy between the content of the designer's or users' original intentions and beliefs and the artifact's actual physicochemical structure and dispositions. The artifact's intended disposition is then not one of its actual dispositions.

Fourth, regardless of whether or not the artifact is malfunctioning, the intentional version of the explanation of how it can fulfill its function supports normative ought-claims about what the artifact and its components ought to be able to do.

To show that this proposal indeed holds water, I demonstrated that a modified version of Wybo Houkes and Pieter Vermaas's ICE-theory of technical functions can be combined with my account of artifact explanations to deliver the goods. The two dovetail nicely and handle the issues about the intentionality and normativity of technical functions and their explanation perfectly. Together, they explicate the full connection between function and structure in the context of explanation.

Finally, let me sketch the general picture that underlies the point of this chapter briefly; I will get back to it in the next chapter when I reassess the Dual Nature intuition. Artifacts, like any material object, have numerous dispositions. How exactly they realize these dispositions is something that can be explained by decomposing the overall behavior into simpler sub-behaviors and by giving the categorical bases of the (sub-)dispositions. None of this requires intentionality. The dispositions of any object can be explained in likewise manner.

The difference with natural material objects is that an artifact also 'has' an intended disposition, which is to say that it is an object that has been made to be able to do specific things. If all is well, this intended disposition is one of its actual dispositions, but sometimes all is not well and artifacts malfunction and do not actually have the disposition for which they were intended. To account for that fact, we do need intentionality, because technical functions are partially constituted by a social and intentional context of human intentions, beliefs, and actions. This context can simply be superimposed on the original explanatory project of accounting for the artifact's dispositions. Knowing for which disposition an artifact was intended certainly enriches our understanding of it, but does not affect the structure of the original explanatory project. It only makes explicit that the disposition being explained happens to be embedded in a use plan and that its manifestation is supposed to contribute to the achievement of a user's practical goal. When this context is put in place, we do acquire the additional possibility of explaining how the artifact is *supposed*

to realize the disposition that is its technical function. This is due to the fact that there are objective facts about how the artifact's designer intended it to be able to show the behavior that is its technical function. These explanations are exactly similar to the original ones, except that they are phrased in terms of the designer's intentions and beliefs; they tell you how the designer intended and believed the artifact to show this behavior.

# Conclusion

## 8.1 Looking Back

We have come a long way since the imagined construction of a garden shed that opened the first chapter. The moral of that story and the starting point for our investigations was that technical artifacts exhibit some sort of duality, i.e., can and must be described both as bare physical objects and in terms of what they are for. To learn more about how these two perspectives, physical and intentional, hang together, we have been looking at the relation between an artifact's function and its structure in two specific contexts in which they are systematically and regularly connected: design and explanation. I think it will be useful to recapitulate the important points so as to get a brief yet comprehensive overview of the topic. After that, in sections 8.2 and 8.3, I will explore two directions in which I think my work can fruitfully be extended.

Besides introducing and specifying the questions and issues that occupied us in these pages, the first two chapters also placed them in context. Heeding a broader call for a more empirically informed philosophy of technology the Dual Nature research program, of which the work for this thesis was part, concentrated on technical artifacts – their functions, design, use, explanation, and social context in particular. Its starting point was the observation that technical artifacts belong equally in two different perspectives. One sees them as simple physical objects and the other takes their for-ness into account by describing them as objects that have been purposefully designed for a specific purpose. Although these perspectives are rather different (the latter presupposes intentionality and allows for normative claims about artifacts, which have no place in the former), they are also somehow intertwined in technical artifacts.

To explore this intertwinement further, I homed in on the relation between an artifact's function and its structure. Function and structure are regularly and systematically brought in contact in two contexts: the design and explanation of technical artifacts. The design context was the focus of chapters 3 and 4, while chapters 5 through 7 revolved around the explanation of artifacts.

In chapter 2 I put forward my basic assumptions about design and technical functions. I assumed technical functions are the behaviors or dispositions for which competent and benign designers or users make or select artifacts.

I further adopted the use plan account of design, which analyzes it as a two-tiered process in which both a use plan and an artifact are devised. I limited myself to the latter activity, in which a desired function is transformed into a detailed description of the physicochemical structure of an artifact that has this function. In the light of the Dual Nature intuition one transition in artifact design is of particular interest: that from an intentionality-presupposing description of desired function to a purely non-intentional description of an artifact's physicochemical structure. This transition has three aspects: it moves (1) from dispositional to categorical, (2) from abstract to concrete, and (3) from intentional to non-intentional or physical. I argued that the first two aspects are what makes design particularly difficult; it is the challenge to find a concrete and categorical description that will realize the abstractly described behavior that is the starting point of a design process. The third aspect boils down to the well-known transition from desiring or intending something and whatever actually comes of it.

To learn more about how designers cope with the challenge, I turned to design methodology in chapter 3. Since that discipline is concerned with what effective design processes look like, I assumed it had to contain clues as to how designers deal with the function-structure transition. The results of this exploration turned out to be ambivalent. To the extent that design methodologists have reflected on the general nature of design, they seem to have been somewhat insensitive to the specific difficulty I identified. The design methodologists I discussed, Herbert Simon, Donald Schön, and John Gero, do not address the issue directly and the answers that can be distilled from their work were either too vague (Schön) or too sweeping (Simon, Gero) to be of real help. This is no reason for despair, however, since design methodologists do give plenty of practical advice to designers in the form of phase models, methods, tools, and techniques. If this advice is endorsed throughout large parts of the design methodology community, it is plausible to assume it is sound, i.e., that the recommended models, methods, tools, and techniques indeed contribute to making the function-structure transition. A closer look at them should therefore reveal something about how the transition from function to structure is accomplished.

In chapter 4 I scrutinized one prominent method: functional decomposition. The point of this method is to break down an overall behavior of an artifact into sub-behaviors. A survey of different treatments of it revealed various features: Functional decomposition specifies an overall behavior in more detail; it determines bits of categorical information of the prospective artifact; it makes the overall behavior more concrete; and it stops when the lowest level of sub-behaviors are considered to be sufficiently basic. The distinction between dispositional and categorical descriptions of artifacts already accounted for a

number of these features. But to analyze functional decomposition further, I tapped into event metaphysics to see what sort of relations can hold between behaviors and sub-behaviors. An inventory of these relations led me to distinguish different possible senses of functional decomposition. Functional decomposition can stand for:

(1) Identifying members of a set of behaviors that are grouped together for some reason such as their contributing to a common goal, spatial or temporal collocation, or being performed by the same (prospective) artifact;
(2) Identifying a behavior's spatial, temporal, or other parts;
(3) Identifying a supervenience base for a behavior (because I distinguished various types of event supervenience, this sense of functional decomposition actually comprises five additional senses).

Functional decomposition in all these senses contributes to the transition from function to structure. It specifies the initial description, breaks down the overall behavior into simpler and more basic sub-behaviors, and gradually reduces the abstractness of the descriptions of behaviors by decomposing them into more concrete behaviors. In the course of the process the designer makes decisions that narrow down the amount of categorical possibilities. First, the sub-behaviors are ordered in space and time. Second, some functional decompositions specify the inputs and outputs of sub-behaviors in great detail, which can strongly suggest particular categorical details and even entail them when a little background knowledge is added. Third, specific ways of breaking down a behavior into sub-behaviors can be highly suggestive of specific implementations, again especially for someone with the relevant background knowledge. In the end, then, a process of functional decomposition does not bridge the gap between function and structure completely, but it does narrow that gap considerably. What remains to be done, however, is figuring out how the identified sub-functions can be grouped together and implemented by suitable physical components.

This transition from function to structure in design has its mirror image in the context of explanation. An artifact can fulfill its function in virtue of its physicochemical structure and that can be explained. Here too, the three aspects need to be carefully distinguished. I suggested that the connection splits in two, with one part being an account of artifact explanation and the other a function theory. In chapters 5 through 7 I explicated this explanatory connection.

Philosophers have thought a lot about explanation, so chapter 5 presented

an overview of the main results of that thinking. We looked at many possible construals of this explanatory relation and I will not repeat them all here. Today's orthodoxy seems to be that a unification and causal accounts of explanation both resonate strongly with scientific explanatory practices. At first sight, unifcationism does better for high-level theoretical explanation, while causalism is more in tune with lower-level explanation of singular events, but both camps have made impressive efforts to accommodate these seeming weaknesses.

Given this state of affairs, I felt free to draw on resources from the causal tradition of explanation for my own work in chapter 6. The causal-mechanical model of explanation seemed to be tailor-made for an account of artifact explanation, since the very thought that gave birth to that model was that the world is rather like a big machine. Examples of artifact explanation showed that they are basically explanations of how artifacts work. The explanandum is the artifact's disposition to show a particular behavior, typically the behavior that is its technical function. The explanans shows how it is that the artifact has the disposition to show this behavior by describing its sub-behaviors and the physical components that take care of these sub-behaviors. More specifically, I identified and discussed two explanatory strategies for explaining an artifact *a*'s behavior *b*, one top-down (TD) and one bottom-up (BU):

> (TD) Describe *b*'s sub-behaviors $b_1$, $b_2$, ..., $b_n$, their sub-behaviors, and so on. Continue until the level of basic sub-behaviors. Include the spatiotemporal configuration of the behaviors. (Put succinctly: give the functional decomposition of *b*.) Point out which structural components take care of the basic sub-behaviors.
> (BU) Describe *a*'s physical components $c_1$, $c_2$, ..., $c_n$, sub-components, and so on. Continue until the level of basic physical components. To the extent that this is possible, give the categorical bases for all of the $c_i$'s relevant dispositions. Then, also give the functional decomposition to show how the lower-level behaviors just accounted for compose into *a*'s overall behavior *b*.

Having explicated these strategies, I went through some lengths to show that they both deliver self-contained explanations. A (TD) style explanation is not inherently incomplete or otherwise wanting. The appropriateness of a (TD) or (BU) explanation depends on context and the specific explanatory question being asked.

The purpose of chapter 7 was to show how this account of artifact explanation ties in with the observation that artifacts belong in an intentional perspective. The basic account did not say much about human intentions or

normativity, while the starting point of this project was that these factors are essential to a proper understanding of technical artifacts. The reason for this is that a function theory should take care of all issues about intentionality and normativity. When artifact explanations are used to explain an artifact's function, they can simply inherit the intentionality and normativity that sticks to functions. They are then drawn into the intentional realm and become explanations of an artifact's intended behavior, or, more precisely, explanations of how an artifact's designer or its users intended and believed the artifact to have the disposition that is its function. For such intentional artifact explanations, we must distinguish two cases. First, when they are given of a well-functioning artifact they coincide fully with how the artifact actually works. The designer's or users' intentions were successful and their beliefs correct, so the intentional version of the explanation forms a transparent window on the artifact's actual physicochemical structure and its dispositions. Second, when they are given of a malfunctioning artifact, there must be at least one discrepancy between what the explanation says about how the designer or users intended the artifact to be and what the artifact's actual structure is. In both cases, however, the intentional artifact explanation of the artifact's function supports normative claims to the effect that this is how the artifact ought to work. This normativity derives from a social division of epistemic labor. Roughly, designers are 'artifact-experts' and their job is to make sure an artifact has the right dispositions for it to be able to function correctly. Non-designing artifact users have a social-epistemic right (which is also institutionalized as a legal right in many contemporary societies) to trust their claims about what an artifact's technical function is and what it ought to do. In the end, then, the complete connection between an artifact's function and its structure in the context of explanation can be fully explicated by a function theory such as the modified ICE-theory combined with my account of artifact explanations.

## 8.2 The Dual Nature of Technical Artifacts Revisited

I recounted at the beginning of chapter 1 that the research for this thesis was inspired by the intuition that there is something special about technical artifacts. While they are plain physical objects, they are also inherently related to human intentions. Hence the Dual Nature intuition:

> (DN) Technical artifacts exhibit some sort of duality, consisting in their equally belonging in an intentional and a physical perspective.

The project of this thesis was to explore ways in which these two perspectives are related, specifically as it manifests itself in the relation between an artifact's

function and its structure. For that reason we have talked quite a bit about intentionality, functions, and physicochemical structure and how these three are connected in design processes and explanations of artifacts. I want to use this section to take a step back and put together an overall picture of this dual nature of technical artifacts that emerges from the foregoing. I will also compare and contrast this picture with some of the initial formulations of the idea.

Since I take it that most people will have a reasonably articulate notion of what non-artificial physical objects are, I want to start with looking at how the intentional perspective on artifacts leads us to understand them as somehow importantly different from non-artificial objects. I think there are two essential factors: Technical artifacts are (1) purposefully made (2) for a specific purpose. That is a cryptic way of saying that technical artifacts are the result of intentional design and creation and that they have technical functions. I will unpack both factors a bit further.

As we saw in chapters 2 through 4, the claim that an artifact is intentionally designed brings along several features. It means that the artifact is embedded in a use plan that is devised along with the artifact itself in a design process. The artifact contributes to the realization of some desired practical goal when it is used in the appropriate way. Doing so is its (initial) technical function even though it may acquire additional functions later by being embedded in different socially accepted use plans. New functions may even replace the initial function. It further means that its behaviors and the components taking care of those behaviors will have been carefully thought out, selected, tested, and put together. Because artifacts are intentionally designed, we can make normative claims about them. In virtue of the social division of epistemic labor, intentional design (especially by professional designers) also gives rise to the strongest type of normative claims about malfunctioning artifacts. Because artifacts are designed for a specific function, they ought to be able to perform that function. It is the social-epistemic and legal responsibility of designers to see to it that they are. These features set technical artifacts apart from natural objects, which have not been intentionally designed – not by human designers at least.

The other component of the intentional perspective is that technical artifacts have a technical function. This, as we saw in chapter 2, is not an intrinsic property of them, but a relational property. It boils down to the artifact standing in a complex relation with (competent and benign) designers or users and their intentions and beliefs and a socially accepted use plan. Having a function presupposes intentional notions such as goals, intentions, beliefs, and a social context of a designer and users. (It is interesting to note that nothing prohibits natural objects from having a technical function. If it so happens that a natural object has the right dispositions to fit in a given use plan, it can simply

be co-opted into a community of users and thereby acquire a technical function. This way a sharp piece of rock can acquire the function of a knife.)

If artifacts have technical functions, we can give normative judgments about them: They ought to be able to perform their function. The precise sense and force of this ought may vary. For initial technical functions it is stronger than for subsequently acquired functions because an artifact's original technical function is what it was designed for and therefore it should definitely be able to perform that function. Its designer is responsible for that. For subsequently acquired functions, the ought is somewhat weaker. Its support is the experience of successful use of a community of users or testimony about this, but this new function is not the artifact's *raison d'être* and there is no division of labor according to which someone was responsible for seeing to it that the artifact can indeed perform this function. We can further compare objects with technical functions to other items with the same functions to evaluate them with regard to how well they perform this function. (Of course, you can compare any two or more objects in arbitrarily many regards, but the having of a technical function provides a more objective ground for comparison and evaluation.)

The physical perspective on artifacts is fairly straightforward. Artifacts are concrete material objects made up of physical components with all manner of spatial and physicochemical properties. So described, they are completely on a par with non-artificial material objects. The only difference would be that some artifacts have fairly 'unnatural' properties compared to what can be found in nature left to its own devices, such as perfect squareness, being made of pure iron, having many interlocking mechanical parts, etc. This, however, is only a gradual difference and it is certainly not a universal hallmark of artifacts.

What, then, could the 'dual nature of technical artifacts' amount to? The first thing to note is that, to the extent that 'nature' hints at an ontological reality, the phrase is slightly misleading. The idea that artifacts essentially consist of two ontologically different types of stuff – some sort of Cartesian dualism for artifacts – is absurd. Artifacts are material objects. They are restlessly composed of physical material and do not have a mental life of their own or anything like it. What does make them special are the two things we just noted, to wit that they are intentionally designed and that they have a technical function. Of these two the former seems to be the most distinctive feature since we also saw that natural objects can acquire technical functions. Still, technical artifacts have a function by definition and that cannot be said of natural objects.

If that is right, then we should say an artifact's first 'nature' is its physicality, its being a concrete material object with everything that is brought along by that. Its second 'nature' consists in its essential connection with designers and their intentions and beliefs and with one or more use plan(s) (and through this with a community of users and their goals and actions). That they

have technical functions follows from this. Technical artifacts have these connections essentially; being a technical artifact requires having them (at least for the types of technical artifacts that are the subject of this thesis). These connections are relational-historical: having been intentionally designed for some purpose is a historical property of artifacts that relates them to designers and their intentions and beliefs. This supports the earlier point that the Dual Nature intuition should not be read as an ontological claim. Because artifacts have this second 'nature' they differ from natural objects, which have no such connections with human intentionality and action. (Of course, their first 'nature' or materiality sets them apart from purely intentional, mental, and social objects such as promises, theories, and marriages.)

The picture this gives is that technical artifacts are material objects that are by definition or essentially embedded an intentional-social context. They are physical objects surrounded by a ring of intentionality and sociality. As a result they have a legitimate and natural place in both a physical and an intentional 'conceptualization of the world,' as the original formulations of the Dual Nature program put it: 'In so far as technical artefacts are physical structures they fit into the physical conception of the world; in so far as they have intentionality-related functions, they fit into the intentional conception. Both conceptualisations are necessary for characterizing technical artefacts' (Kroes and Meijers 2006: 2). The claim that these two conceptualizations are 'fundamentally different ways of viewing our world' (ibid.: 2), however, should not be misread as a suggestion that the physical and intentional conceptualizations are entirely separate, which, in turn, might lead to a picture of artifacts that is more dualistic than warranted. What I just said about the dual nature of technical artifacts is more in line with Maarten Franssen's recent remarks about the two conceptualizations of the world. He sees the intentional perspective or conceptualization as an extension of the physical one. Everything that can be said in the physical conceptualization can also be said in the intentional conceptualization, but the latter adds intentional and intentionality-presupposing notions to the former. Says Franssen: 'The physical conceptualization must be seen as being contained in the intentional conceptualization, or the intentional description as being hooked onto the physical description. In the intentional 'conceptualization of the world', if we are to retain for a moment this terminology, the physical description of the world is presupposed' (Franssen 2007).

The fact the intentional conceptualization presupposes the physical one does not imply that everything said in the terminology of the intentional conceptualization will contain physical terms. Artifacts are often described teleologically, in terms of their technical functions, or the purposes for which they were made or currently used. These are clearly descriptions that

presuppose intentionality and therefore belong in the intentional perspective. At the same time, however, they may well contain elements from the physical perspective. An artifact's function is the disposition its designers intended it to have, but when we describe this disposition in isolation, without mentioning that it is intended and embedded in a use plan, we end up with a purely non-intentional physical description of what the artifact can do.

We saw that design processes start with intentional descriptions of an artifact's desired behavior. The point of a design process is to devise a blueprint for an artifact that will actually be able to show this desired behavior. A designer accomplishes this task by putting her beliefs to work – beliefs about how behaviors can be decomposed into sub-behaviors, about which materials and components have the right dispositions to jointly take care of these sub-behaviors, etc. When the described artifact is actually created – i.e., when the designer's intentions and beliefs are confronted with the non-intentional physical reality – the artifact's 'physicality' is its first nature, while its second nature consists in its being an object that was designed and created for its (actual or non-actual) ability to perform the specific behavior that is now its function. That is basically how the dual nature plays out in the design process.

Something that is more or less the reverse of this happens in the context of explanation; that is what I argued in chapter 7. First, the intended behavior that is an artifact's technical function is identified by means of a function theory. This function theory relates the intended behavior with the artifact's actual structure through the intentions and beliefs of designers or later users. We might say that this provides a characterization of the intentional ring that surrounds the artifact as physical object. Artifact explanations, as described in chapter 6, can then be used to explain how the artifact is supposed to be able to show this intended behavior. These explanations, which are by themselves purely physical, are drawn into the intentional perspective. They describe how the artifact's designers or users intended or believed it to have the disposition that was identified as its function and so explicate how the artifact's second intentional nature is supposed to be related to its first physical nature. When the artifact functions as intended, the intentionality of the explanation can easily be stripped away since it consists merely of successfully carried out intentions and true beliefs about the artifact that all correspond with how the artifact really is. When the artifact malfunctions, however, the intentional version of the explanation is not completely true of the actual artifact and there remains a discrepancy between the artifact-as-intended and the actual artifact.

This is how I think the Dual Nature intuition can be made more precise. To conclude this thesis, I want to take a look ahead at possibilities for further research.

## 8.3 Artifact Kinds and Other Suggestions for Further Research

I will explore one topic in some detail and hint at a few others. Something that lingered in the background but that I have not had the time to go into is the notion of artifact kinds. Thinking about artifacts as coming in kinds is very common. A chair is a different kind of thing than a bridge and a car is not a TV. Artifact kinds are also ordered hierarchically; the kind stool splits up in sub-kinds of three-legged and four-legged stools. Watches can be digital or analog.

In virtue of what do we categorize artifacts in kinds? What is it that makes an artifact a member of one kind as opposed to another? An intuitive commonsense view is that artifact kinds are functional kinds; an artifact is a $\phi$-type artifact because it has the function to $\phi$. In fact, this is not only a commonsense view, it is also the view espoused by respectable philosophers. David Wiggins subscribes to it when he says: 'A clock is any time-keeping device, a pen is any rigid ink-applying writing implement, and so on' (Wiggins 2001: 87). Stephen Mumford also seems to endorse it: 'What it is that makes certain artefacts the thing that they are is that they have a particular set of dispositions. ... Anything which has this disposition ... is a thermostat regardless of the constitution that affords such an ability' (Mumford 1998: 8–9). Basically the view seems to be that, for any artifact $x$, $x$ is an artifact of the kind $K_\phi$ iff $x$ has the function to $\phi$, where $K_\phi$ is the functional kind constituted by things having $\phi$ as their function. This view of artifact kinds – let's call it the standard view for the sake of brevity – works pretty well for a lot of cases. It is true that a thing with the function to light a room is a lamp, just as it is true that a pencil is for writing with.

In spite of its intuitive appeal, I think the standard view oversimplifies the notion of artifact kinds. Chapter 7 provides material to argue this point. In particular, I want to make two points that cast doubt on the standard view; one deriving from considerations about artifact explanation and the other from the normativity that adheres to claims about an artifact's technical function(s). For both points we will need the assumption that technical artifacts can change their technical function or can at least acquire new technical functions over the course of their lifetime. Support for this assumption can be derived from the ICE-theory of technical functions. If an artifact is embedded in a different socially accepted use plan, it changes its technical function, according to that theory. The assumption is also independently supported by everyday experience. Some artifacts do indeed change their function. Aspirin is a pain-killer, but it now also has the function of preventing blood clots in cardiac patients. Nobody uses an old iron to iron their clothes anymore, but I have seen quite a few people that use one as a doorstop. Old car tires are often used to create garden swings. Although these simple examples do not prove unequivocally that artifacts can change their function, saying that they do seems to be a very natural of describing what happens.

Here is my first point. Artifact explanation, as we saw in chapter 6, proceeds by showing how an artifact's behavior is realized by various sub-behaviors, which are made possible by the underlying categorical properties of the artifact. Artifact explanations of how an artifact performs its original function are governed by a principle that I did not make explicit in chapter 6 and 7, but which is readily seen to be correct. We could dub it the principle of Maximal Component Utilization (MCU):

> (MCU) Almost all of an artifact's components and many of their specific features figure directly or indirectly in an explanation the artifact's functioning.[98]

Explanations of how an artifact fulfills its original function will mention almost all of the artifact's components and many of their specific features. If some component is not mentioned in the explanation of how the artifact fulfills its function, there will often be a follow-up explanation of how this component or feature is required for assembly or maintenance.

The reason why this principle holds is the way engineering design works. Engineering design is subject to various constraints of effectiveness, efficiency, and rationality. Designers have to be effective at meeting the requirements set out for them by customers, i.e. they have to design a product that actually does what it is supposed to do and does it quickly, cheaply, elegantly, or differently. They also have to do this within a reasonable amount of time and with limited amounts of money and materials. Furthermore, designers obviously base their design decisions on broadly rational considerations. As a result, the artifact's components, including many of their specific details and configuration, will be the way they are for specific reasons. They should contribute directly or indirectly to the function(s) of the artifact, stimulate and enable convenient or proper use and maintenance, or facilitate some part of the artifact's assembly process. Of course, some details will be unintended and inevitable consequences of other intended details and some will have to do with aesthetic considerations, but the great majority will be purposefully thought out. Modern-day engineering design leaves little room for frivolities. Because of this, normal functioning of an artifact requires that almost all of its components and their features be in place and functioning properly. Hence (MCU).

Now think of a case in which an artifact changes its function. The iron used as doorstop for example. An explanation of how it fulfills this function, while

[98] As Tim Lewens pointed out to me, if this principle is to be fully general, the dispositions of redundant components and subsystems that kick in if the primary system fails should be understood to be part of the functioning of an artifact.

admittedly rather trivial, will embody a violation of (MCU). Nothing will be said about the iron's capacity to retain heat and the iron's specific shape and its handle also have little to contribute to the stopping of doors. This phenomenon drives a conceptual wedge between an artifact's original function and its later function(s). In some sense, an artifact's original function sticks to it in a way that later functions do not. I think the explanation for this is that only an artifact's original function is closely related to the artifact's kind, while later acquired functions are not. Artifacts may change their technical functions, but they do not change their kind – or so I submit. The doorstop-iron still *is* an iron in an important sense, and not a doorstop. Imagine someone insisting that the doorstop-iron is really a doorstop and not an iron. I think we would find this insistence rather odd. This is explained neatly if we accept that the sense in which the doorstop-iron still is an iron, is that it still is a member of the artifact kind iron.

If this is correct, we have an argument against the standard view. That something has a technical function to $\phi$ does not make it a member of the kind of artifacts that typically have $\phi$ as their function. An artifact's technical function is susceptible to change; its kind is not. Therefore the standard view that couples an artifact's technical function to its kind cannot be correct. Another way to put the same point is that artifact kinds do not seem to be mere functional kinds, the membership of which requires nothing more than the having of a function.

The second point is based on the observation that normative claims about an artifact's original technical function are stronger than those about its new technical function(s). That this is so is easily appreciated by imagining someone who wants to complain about an artifact not performing an acquired function well. No producer of car tires in his right mind will take it as a serious defect of his tires that they do not make for comfortable garden swing seats. The sense in which a car tire ought to perform its original function of, say, supporting cars and enabling smooth steering and braking, is much stronger than that in which it ought to make a good swing seat. The latter claim is based on no more than the past experience of a community of people who were looking for a cheap way to build a garden swing, but the former is based on what professional designers designed and created car tires for. The original function of the tire is the reason it exists. In virtue of the social-epistemic division of labor between expert designers and lay-users, users of tires can trust designers to see to it that a tire can fulfill its original function well. This division of labor underlies the stronger normative sense that attaches to claims about an artifact's original function.

In view of the tight connection between an artifact's original function and its kind, I think the best way to make sense of this is to say that, in fact, an

artifact's kind licenses the particular strong type of normative claims about what an artifact ought to be able to do. Because an artifact is a member of the kind car tire, it ought to be able to function as one (barring worn-out or broken exemplars). We have a social-epistemic and even a legal right to expect this.

That provides a second argument against the standard view. Kind membership supports stronger normative claims than the mere having of a technical function. This would be odd if the having of a particular technical function is all there is to being a member of a particular artifact kind. So again, we see that having the technical function to $\phi$ and being a member of the kind of artifacts that typically have $\phi$ as their function come apart and that provides another reason for thinking that the best construal of artifact kinds may not be that they are functional kinds constituted by everything that has a particular technical function.

These two points make it plausible that having a technical function to $\phi$ and being a member of $K_\phi$ are not coextensive, as the standard view has it. So what next? I see two solutions. The first is to modify the standard view to incorporate the superior importance of an artifact's original function. The view would then be that, for any artifact $x$, $x$ is an artifact of the kind $K_\phi$ iff $x$ originally had the function to $\phi$, where $K_\phi$ is the historical-functional kind constituted by things having $\phi$ as their original function. Yet another option makes the link with functions even weaker by construing artifact kind membership as being determined by a maker's successful intention to create an artifact of a certain kind $K$. On this construal, artifact kinds are intentional-historical kinds.

While the former view may seem to be the most appealing at first sight because it deviates minimally from the standard view, psychological research on artifact concepts strongly suggests that this view, too, does not capture the way people think about artifact kinds adequately. Malt and Johnson (1992) show that having a particular original function is neither necessary nor sufficient for being considered a member of the kind of artifacts that typically have that function. It may be possible to contest their interpretation of the empirical findings, but this is not the place to do so. It suffices to take note of their conclusion and accept that it provides reason to look for yet another account of artifact kinds.

Amie Thomasson promotes the intentional-historical view (Thomasson 2003), crediting Paul Bloom (1996) and Jerrold Levinson (1979; 1989; 1993) for inspiration. On this view what kind of thing an artifact is, is determined by its maker's successful intention to create something of a particular kind. Built into the notion of a successful intention is that the maker has to have a largely correct substantive concept of the nature of the artifact she intends to create. Correctness of this concept is settled by correspondence with the concept that previous makers had in mind when creating artifacts of the relevant kind. The

successfulness of the intention consists in the maker's succeeding, by and large, to impose the kind-relevant features that form the content of her 'largely correct substantive concept' on a particular object. Thomasson says nothing specific about what these kind-relevant features may be, but it is reasonable to assume that they will be heavily influenced by how the artifact is supposed to carry out its (original) technical function. That is how the link with original proper functions is forged on this view.

This topic merits further attention. Both alternatives to the standard view can be developed further. The historical-functional kinds view has to account for the psychological research that indicates that original function is not what determines people's use of artifact kind concepts. And it has to deal with the objection that designers can bestow atypical original functions on token artifacts of a kind. Thomasson gives the example of a show-off boat that is never meant to actually sail but only to impress other people (2003: 594). The intentional-historical view needs to say more about the exact nature of 'kind-relevant features' and it is incomplete without an account of how new kinds are inaugurated. No matter which view turns out to be most promising, it is also important to have a complementary account of artifact kind individuation. What sort of criteria determine whether one kind differs from another? How broad or narrow are artifact kinds? Finally, what kind of kinds are artifact kinds? Are they mere nominal kinds (as Schwartz (1978; 1980) and Wiggins (2001) argue) or are they metaphysically more robust (as Putnam (1975), Elder (1989), and Thomasson (2003) hold, each in their own way)?

Artifact kinds is not the only thing that needs more thinking. So here are some other topics that I have come across while working on this thesis but did not have the time to delve into.

First, I have only looked at one design methodological tool in detail, functional decomposition. To my mind, this has been a rewarding undertaking; a close look at functional decomposition provided much more clarity and yielded important insights into how this method helps to make the transition from function to structure. The field of design methodology has developed numerous methods, tools, and techniques that deserve similar attention. To mention but a few of well-known examples: *Quality Function Deployment* (QFD) is a method to systematically map customer requirements onto product functions and characteristics. *Morphological analysis* tries to generate alternatives by systematically studying and comparing the shapes or forms a product might take. A method that is specifically about the redesign and optimization of existing products, is the *Value Engineering Method*. It identifies the costs and functional values of an artifact's components and then tries to increase the difference between these two by reducing the costs or increasing the value.

These methods can be found in many design methodology handbooks; Ullman (1992), Roozenburg and Eekels (1995), Cross (2000), and Eggert (2005) are just a few examples.

Second, it is probably a platitude that design inevitably involves sparks of creativity (whatever that means exactly). The amount and intensity of these sparks may vary from one project to the next and one may debate the relative importance of rational-analytical thinking and creativity. I have focused exclusively on rational systematic methods for design and done nothing to clarify the role of creativity. I think creativity in design, too, is something that needs to be examined in more detail, especially since writers on creativity and creative methods tend to be fairly vague and ambiguous (which may of course be due to the inherent vagueness and ambiguity of their topic). I have to confess that I do not know what the best starting point would be for such an investigation, but one option might be to turn to design methodology again and study some of the various creativity-enhancing methods it recommends.

Third, I have studied the transition from desired function to structure in design processes, but design also involves a transition from a client's needs or goals to a desired function. On the use plan account of design discussed in chapter 2, this transition is made in the design of a use plan for an artifact. This process, in which a series of actions is devised that will realize a goal, should be mapped out further. The relevance of this goes far beyond the philosophy of technology, since the ability of people to construct plans to realize their goals is a topic right in the heart of general action theory. It has to do with practical rationality and deliberation, means-ends reasoning, etc. Seeing how this plays out in engineering design, which is supposed to be a systematic way of creating use plans, may not only improve or understanding of engineering design, but may also shed light on general issues in action theory.

Fourth and finally, there are lots of fascinating similarities and dissimilarities between technology and biology. For example, the talk of design, evolution of organisms and products, and the use of functions and other teleological notions provide rich fertile grounds of comparison between the two domains. Of course analogies have long been noted and a lot of work has already been done that exploits the (seeming) similarities and dissimilarities. Various authors have proposed evolutionary accounts of technological development (Basalla 1988; Petroski 1992; Dennett 1995; Blackmore 1999; Ziman 2000; Aunger 2000, 2002; Mokyr 2002). Many of the prominent writers on biological function theories have attempted to apply their theory to the products of technology (Neander 1991b; Millikan 1999). And Tim Lewens has worked on the notion of design in biology and technology (Lewens 2004). Given the traditional focus of philosophers of technology on the ethical and societal effects of technology, it is hardly surprising that most of the input in those debates

has come from philosophers of biology who, I regret to say, are not always impressively well-informed about technology. It is about time philosophers of technology rise to the occasion and bring their specific expertise to bear on these debates. Doing so should not only enliven the debate, but also serve to correct any naïveté about technology and its products. Perhaps this thesis can provide something of a bridgehead for this undertaking.

# References


Achinstein, Peter (1983), *The Nature of Explanation*. New York: Oxford University Press.

Armstrong, David (1968), *A Materialst Theory of the Mind*. London: Routledge.

Asimow, M. (1962), *Introduction to Design*. Englewood Cliffs, NJ: Prentice Hall.

Aunger, Robert A. (2002), *The Electric Meme: A New Theory of How We Think*. New York: The Free Press.

–, ed. (2000), *Darwinizing Culture: The Status of Memetics as a Science*. Oxford: Oxford University Press.

Basalla, George (1988), *The Evolution of Technology*. New York: Cambridge University Press.

Bechtel, William, and Robert C. Richardson (1993), *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Bijker, Wiebe E. (1995), *Of Bicycles, Bakelites and Bulbs: Toward a Theory of Sociotechnical Change*. Cambridge, MA: MIT Press.

Bijker, Wiebe E., Thomas Hughes, and Trevor Pinch, eds. (1987), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: MIT Press.

Blackmore, Susan J. (1999), *The Meme Machine*. Oxford: Oxford University Press.

Bloom, Paul (1996), 'Intention, History, and Artifact Concepts', *Cognition* 60(1): 1–29.

Bromberger, Sylvain (1966), 'Why-Questions', in R.G. Colodny (ed.), *Mind and Cosmos*, Pittsburgh: University of Pittsburgh Press.

Bucciarelli, Louis L. (1996), *Designing Engineers*. Cambridge, MA: MIT Press.

Cartwright, Nancy (1983), *How the Laws of Physics Lie*. Oxford: Clarendon Press.

– (1999), *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Casati, Roberto, and Achille C. Varzi, eds. (1996), *Events*. Aldershot: Dartmouth.

Chandrasekaran, Balakrishnan (2005), 'Representing Function: Relating Functional Representation and Functional Modelling Streams', *Artificial Intelligence for Engineering Design, Analysis, and Manufacturing* 19: 65–74.

Chandrasekaran, Balakrishnan, and John R. Josephson (2000), 'Function in Device Representation', *Engineering with Computers* 16(3–4): 162–177.

Craver, Carl F. (2001), 'Role Functions, Mechanisms, and Hierarchy', *Philosophy of Science* 68(1): 53–74.
Cross, Nigel G. (2000), *Engineering Design Methods: Strategies for Product Design*. 3rd ed. Chichester, etc.: John Wiley & Sons.
–, ed. (1984), *Developments in Design Methodology*. Chichester: John Wiley & Sons.
Cummins, Robert (1975), 'Functional Analysis', *Journal of Philosophy* 72(20): 741–765.
Davidson, Donald (1980), *Essays on Actions and Events*. Oxford: Oxford University Press.
De Regt, Henk W., and Dennis Dieks (2005), 'A Contextual Approach to Scientific Understanding', *Synthese* 144(1): 137–170.
De Ridder, Jeroen (2006), 'Mechanistic Artefact Explanation', *Studies in History and Philosophy of Science* 37(1): 81–96.
– (2007), 'The (Alleged) Inherent Normativity of Technological Explanations', *Technè*.
Dennett, Daniel C. (1978), *Brainstorms*. Cambridge, MA: Bradford Books.
– (1995), *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.
Descartes, René (1999), *Principia Philosophiae*. Translated by George MacDonald Ross. Original edition, 1644.
Dietz, Jan G. (2006), *Enterprise Engineering*. Delft: Delft University of Technology Manuscript.
Dorst, Kees (1997), *Describing Design: A Comparison of Paradigms*. Ph.D. Thesis. Faculty of Industrial Design, Delft: Delft University of Technology.
Dorst, Kees, and Pieter E. Vermaas (2005), 'John Gero's Function-Behaviour-Structure Model of Designing: A Critical Analysis', *Research in Engineering Design* 16(1/2): 17–26.
Dupré, John (1993), *The Disorder of Things*. Cambridge, MA: Harvard University Press.
Eggert, Rudolph J. (2005), *Engineering Design*. Upper Saddle River, NJ: Pearson Education.
Elder, Crawford L. (1989), 'Realism, Naturalism, and Culturally Generated Kinds', *The Philosophical Quarterly* 39(157): 425–444.
Ferré, Frederick (1988), *Philosophy of Technology*. Englewood Cliffs, NJ: Prentice Hall.
Fodor, Jerry A. (1974), 'Special Sciences (Or: The Disunity of Science as a Working Hypothesis)', *Synthese* 28: 97–115.
– (1997), 'Special Sciences: Still Autonomous after All These Years', *Philosophical Perspectives* 11: 149–163.

Franssen, Maarten (2007), 'Design, Use, and the Physical and Intentional Aspects of Technical Artifacts', in Pieter E. Vermaas (ed.), *Spt '05 Proceedings*, Delft.

French, Michael J. (1985), *Conceptual Design for Engineers*. London: Design Council.

Friedman, Michael (1974), 'Explanation and Scientific Understanding', *Journal of Philosophy* 71(1): 5–19.

Garfinkel, A. (1981), *Forms of Explanation: Rethinking the Questions of Social Theory*. New Haven, CT: Yale University Press.

Gero, John S. (1990), 'Design Prototypes: A Knowledge Representation Schema for Design', *AI Magazine* 11(4): 26–36.

– (2002), 'Categorising Technological Knowledge from a Design Methodological Perspective', Paper read at Technological Knowledge: Philosophical Reflections, June 13–15, 2002, at Boxmeer, Netherlands.

Gero, John S., and U. Kannengieser (2002), 'The Situated Function-Behaviour-Structure Framework', in John S. Gero (ed.), *Artificial Intelligence in Design '02*, Dordrecht: Kluwer, 89–104.

Gero, John S., and Michael A. Rosenman (1990), 'A Conceptual Framework for Knowledge-Based Design Research at Sydney University's Design Computing Unit', *Artificial Intelligence in Engineering* 5(2): 65–77.

Gero, John S., K.W. Tham, and H.S. Lee (1992), 'Behavior: A Link between Function and Structure in Design', in D.C. Brown, M.B. Waldron and H. Yoshikawa (eds.), *Intelligent Computer Aided Design*, Amsterdam: Elsevier, 193–225.

Glennan, Stuart S. (1996), 'Mechanisms and the Nature of Causation', *Erkenntnis* 44(1): 49–71.

– (2002), 'Rethinking Mechanistic Explanation', *Philosophy of Science* 69: S342–S353.

Glymour, Clark (1982), 'Causal Inference and Causal Explanation', in R. McLaughlin (ed.), *What? Where? When? Why?*, Dordrecht: Reidel.

Goldman, Alvin (1970), *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Hampton, Jean (1994), 'The Failure of Expected-Utility Theory as a Theory of Reason', *Economics and Philosophy* 10(2): 195–242.

Hempel, Carl G. (1962), 'Explanation in Science and History', in R.G. Colodny (ed.), *Frontiers of Science and Philosophy*, Pittsburgh: University of Pittsburgh Press.

– (1965a), 'Aspects of Scientific Explanation', in, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: The Free Press, 331–496.

– (1965b), 'The Logic of Functional Analysis', in, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: The Free Press, 297–330.

Hempel, Carl G., and Paul Oppenheim (1948), 'Studies in the Logic of Explanation', *Philosophy of Science* 15: 567–579.
Hills, Richard L. (1989), *Power from Steam: A History of the Stationary Steam Engine*. Cambridge: Cambridge University Press.
Hirtz, Julie, Robert B. Stone, Daniel A. McAdams, Simon Szykman, and Kristin L. Wood (2002), 'A Functional Basis for Engineering Design: Reconciling and Evolving Previous Efforts', *Research in Engineering Design* 13: 65–82.
Hitchcock, Christopher R. (1995), 'Discussion: Salmon on Explanatory Relevance', *Philosophy of Science* 62: 304–320.
Houkes, Wybo N. (2006), 'Knowledge of Artefact Functions', *Studies in History and Philosophy of Science* 37(1): 102–113.
Houkes, Wybo N., and Pieter E. Vermaas (2004), 'Actions Versus Functions: A Plea for an Alternative Metaphysics of Artifacts', *The Monist* 87(1): 52–71.
– (2006), *Artefacts: From Functions to Plans of Use*. Delft: Delft University of Technology Manuscript.
Houkes, Wybo N., Pieter E. Vermaas, Kees Dorst, and Marc J. De Vries (2002), 'Design and Use as Plans: An Action-Theoretical Account', *Design Studies* 23(3): 303–320.
Hubka, V., and W.E. Eder (1988), *Theory of Technical Systems: A Total Concept Theory for Engineering Design*. Berlin: Springer Verlag.
Humphreys, Paul (1989), *The Chances of Explanation*. Princeton, NJ: Princeton University Press.
Huygens, Christiaan (1945), *Treatise on Light*. Translated by Silvanus P. Thompson. Chicago: University of Chicago Press. Original edition, 1912.
Ingold, Tim (2000), 'On Weaving a Basket', in Tim Ingold (ed.), *The Perception of the Environment: Essays on Livelihood, Dwelling, and Skill*, London & New York: Routledge.
Jackson, Frank (1998), *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. New York: Oxford University Press.
Jackson, Frank, and Philip Pettit (1990), 'Program Explanation: A General Perspective', *Analysis* 50(2): 107–117.
– (1992a), 'In Defense of Explanatory Ecumenism', *Economics and Philosophy* 8(1): 1–21.
– (1992b), 'Structural Explanation in Social Theory', in David Charles and Kathleen Lennon (eds.), *Reduction, Explanation, and Realism*, Oxford: Clarendon Press, 97–131.
Jeffrey, Richard C. (1971), 'Statistical Explanation vs. Statistical Inference', in Wesley C. Salmon, Richard C. Jeffrey and James G. Greeno (eds.), *Statistical Explanation and Statistical Relevance*, Pittsburgh: University of Pittburgh Press.
Jones, J. Christopher (1970), *Design Methods: Seeds of Human Futures*. London: Wiley.

Jones, J. Christopher, and D.G. Thornley, eds. (1963), *Conference on Design Methods. Papers Presented at the Conference on Systematic and Intuitive Methods in Engineering, Industrial Design, Architecture and Communications*. Oxford: Pergamon Press.

Jones, T. (1997), 'Unification, Reduction, and Non-Ideal Explanations', *Synthese* 112: 75–96.

Kargon, Robert, and Peter Achinstein, eds. (1987), *Kelvin's Baltimore Lectures and Modern Theoretical Physics*. Cambridge, MA: MIT Press.

Kemeny, John G., and Paul Oppenheim (1956), 'On Reduction', *Philosophical Studies 7*: 6–19.

Kim, Jaegwon (1974), 'Noncausal Connections', *Noûs* 8: 41–52.

– (1992), 'Multiple Realization and the Metaphysics of Reduction', *Philosophy and Phenomenological Research* 52(1): 1–26.

Kitcher, Philip (1981), 'Explanatory Unification', *Philosophy of Science* 48: 507–531.

– (1984), '1953 and All That. A Tale of Two Sciences', *The Philosophical Review* 93(3): 335–373.

– (1985), 'Two Approaches to Explanation', *Journal of Philosophy* 82(11): 632–639.

– (1989), 'Explanatory Unification and the Causal Structure of the World', in Philip Kitcher and Wesley C. Salmon (eds.), *Scientific Explanation, Minnesota Studies in the Philosophy of Science, Volume 13*, Minneapolis: University of Minnesota Press, 410–505.

Kitcher, Philip, and Wesley C. Salmon (1987), 'Van Fraassen on Explanation', *Journal of Philosophy* 84: 315–330.

Kroes, Peter A. (1998), 'Technological Explanations: The Relation between Structure and Function of Technological Objects', *Technè* 3(3): 18–34.

– (2000), 'Engineering Design and the Empirical Turn in the Philosophy of Technology', in Peter A. Kroes and Anthonie W.M. Meijers (eds.), *The Empirical Turn in the Philosophy of Technology*, Amsterdam: JAI (Elsevier Science), 19–43.

– (2002), 'Design Methodology and the Nature of Technical Artefacts', *Design Studies* 23(3): 287–302.

– (2006), 'Coherence of Structural and Functional Descriptions of Technical Artefacts', *Studies in History and Philosophy of Science* 37(1): 137–151.

Kroes, Peter A., and Anthonie W.M. Meijers (2000), 'Introduction: A Discipline in Search of Its Identity', in Peter A. Kroes and Anthonie W.M. Meijers (eds.), *The Empirical Turn in the Philosophy of Technology*, Amsterdam: JAI (Elsevier Science), xvii–xxxv.

– (2002), 'The Dual Nature of Technical Artifacts – Presentation of a New Research Programme', *Technè* 6(2): 4–8.

– (2006), 'Introduction: The Dual Nature of Technical Artefacts', *Studies in History and Philosophy of Science* 37(1): 1–4.

Kyburg, H.E. (1965), 'Comment', *Philosophy of Science* 32: 147–151.

Levinson, Jerrold (1979), 'Defining Art Historically', *British Journal of Aesthetics* 19: 232–250.
– (1989), 'Refining Art Historically', *Journal of Aesthetics and Art Criticism* 47: 21–33.
– (1993), 'Extending Art Historically', *Journal of Aesthetics and Art Criticism* 51: 411–423.
Lewens, Tim (2004), *Organisms and Artifacts: Design in Nature and Elsewhere*. Cambridge, MA: MIT Press.
Lewis, David (1973), 'Causation', *Journal of Philosophy* 70(17): 556–567.
– (1986), 'Causal Explanation', in, *Philosophical Papers, Volume 2*, New York: Oxford University Press.
– (1997), 'Finkish Dispositions', *The Philosophical Quarterly* 47(187): 143–158.
Lipton, Peter (1991), *Inference to the Best Explanation*. New York: Routledge.
Lycan, William G. (1999), 'The Continuity of Levels of Nature', in William G. Lycan (ed.), *Mind and Cognition*, Oxford: Blackwell, 49–63.
Machamer, Peter, Lindley Darden, and Carl F. Craver (2000), 'Thinking About Mechanisms', *Philosophy of Science* 67(1): 1–25.
Mackie, John L. (1974), *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon Press.
Malt, Barbara C., and Eric C. Johnson (1992), 'Do Artifact Concepts Have Cores?' *Journal of Memory and Language* 31(2): 195–217.
McLaughlin, Peter (2006), 'Discussion: Mechanical Philosophy and Artefact Explanation', *Studies in History and Philosophy of Science* 37(1): 97–101.
Meijers, Anthonie W.M. (2000), 'The Relational Ontology of Technical Artifacts', in Peter A. Kroes and Anthonie W.M. Meijers (eds.), *The Empirical Turn in the Philosophy of Technology*, Amsterdam: JAI (Elsevier Science), 81–96.
Mellor, D.H. (1974), 'In Defense of Dispositions', *The Philosophical Review* 83(2): 157–181.
– (1982), 'Counting Corners Correctly', *Analysis* 42(2): 96–97.
Menzies, Peter (1988), 'Against Causal Reductionism', *Mind* 97(388): 551–574.
Millikan, Ruth (1984), *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.
– (1993), *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.
– (1999), 'Wings, Spoons, Pills and Quills: A Pluralist Theory of Function', *Journal of Philosophy* 96(4): 192–206.
Mitcham, Carl (1994), *Thinking through Technology: The Path between Engineering and Philosophy*. Chicago: University of Chicago Press.
Mokyr, Joel (2002), *The Gifts of Athena: Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.
Morrison, Margaret (2000), *Unifying Scientific Theories : Physical Concepts and Mathematical Structures*. Cambridge: Cambridge University Press.

Mumford, Stephen (1998), *Dispositions*. New York: Oxford University Press.
– (2006), 'The Ungrounded Argument', *Synthese* 149(3): 471–489.
Nagel, Ernest (1961), *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace and World.
Neander, Karen (1991a), 'Functions as Selected Effects: The Conceptual Analyst's Defense', *Philosophy of Science* 58(2): 168–184.
– (1991b), 'The Teleological Notion of "Function"', *Australasian Journal of Philosophy* 69(4): 454–468.
Oppenheim, Paul, and Hilary Putnam (1958), 'Unity of Science as a Working Hypothesis', in Herbert Feigl, Michael Scriven and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science, Vol. 2*, Minneapolis, MN: University of Minnesota Press, 3–36.
Pahl, G., and W. Beitz (1996), *Engineering Design: A Systematic Approach*. 2nd ed. London: Springer Verlag. Original edition, 1984.
Petroski, Henry (1992), *The Evolution of Useful Things*. New York: Alfred A. Knopf.
– (1994), *Design Paradigms: Case Histories of Error and Judgment in Engineering*. Cambridge: Cambridge University Press.
– (1996), *Invention by Design*. Cambridge, MA: Harvard University Press.
Pettit, Philip (1996), 'Functional Explanation and Virtual Selection', *British Journal for the Philosophy of Science* 47(2): 291–302.
Pinch, Trevor, and Nelly Oudshoorn, eds. (2003), *How Users Matter: The Co-Construction of Users and Technologies*. Cambridge, MA: MIT Press.
Pitt, Joseph (1999), *Thinking About Technology: Foundations of the Philosophy of Technology*. New York: Seven Bridges Press.
Polanyi, Michael (1958), *Personal Knowledge. Towards a Post-Critical Philosophy*. Chicago: University of Chicago Press.
Preston, Beth (1998a), 'Cognition and Tool Use', *Mind and Language* 13: 513–547.
– (1998b), 'Why Is a Wing Like a Spoon? A Pluralist Theory of Function', *Journal of Philosophy* 95(5): 215–254.
Prior, Elizabeth W. (1985), *Dispositions*. Aberdeen: Aberdeen University Press.
Prior, Elizabeth W., Robert Pargetter, and Frank Jackson (1982), 'Three Theses About Dispositions', *American Philosophical Quarterly* 19(3): 251–257.
Putnam, Hilary (1975), *Mind, Language, and Reality*. Cambridge: Cambridge University Press.
Quine, Willard Van Orman (1951), 'Two Dogmas of Empiricism', *The Philosophical Review* 60(1): 20–43.
Railton, Peter (1978), 'A Deductive-Nomological Model of Probabilistic Explanation', *Philosophy of Science* 45(2): 206–226.
– (1981), 'Probability, Explanation, and Information', *Synthese* 48: 233–256.

Rittel, Horst W.J., and Melvin M. Webber (1984), 'Dilemmas in a General Theory of Planning', in Nigel G. Cross (ed.), *Developments in Design Methodology*, Chichester: John Wiley & Sons.
Roozenburg, N.F.M., and J. Eekels (1995), *Product Design: Fundamentals and Methods*. New York: John Wiley & Sons.
Rosenman, Michael A., and John S. Gero (1998), 'Purpose and Function in Design: From the Socio-Cultural to the Techno-Physical', *Design Studies* 19(2): 161–186.
Ruben, David-Hillel (1993), 'Introduction', in David-Hillel Ruben (ed.), *Explanation*, New York: Oxford University Press.
– (2003), *Action and Its Explanation*. Oxford: Oxford University Press.
Ryle, Gilbert (1990), *The Concept of Mind*. London: Penguin Books. Original edition, 1949.
Salmon, Wesley C. (1971), 'Statistical Explanation', in Wesley C. Salmon, Richard C. Jeffrey and James G. Greeno (eds.), *Statistical Explanation and Statistical Relevance*, Pittburgh: University of Pittsburgh Press.
– (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
– (1989), 'Four Decades of Scientific Explanation', in Philip Kitcher and Wesley C. Salmon (eds.), *Scientific Explanation, Minnesota Studies in the Philosophy of Science, Volume 13*, Minneapolis: University of Minnesota Press, 3–219.
Scheele, Marcel (2005), *The Proper Use of Artefacts: A Philosophical Analysis of the Social Constitution of Artefact Functions*. Department of Philosophy, Delft: Delft University of Technology.
Schön, Donald A. (1983), *The Reflective Practitioner.* New York: Basic Books.
– (1984), 'Problems, Frames and Perspectives on Designing', *Design Studies* 5(3): 132–136.
– (1987), *Educating the Reflective Practitioner.* New York: Basic Books.
– (1988), 'Designing: Rules, Types and Words', *Design Studies* 9(3): 181–190.
– (1992), 'Desiging as Reflective Conversation with the Materials of a Design Situation', *Research in Engineering Design* 3(3): 131–147.
Schön, Donald A., and Glenn Wiggins (1992), 'Kinds of Seeing and Their Functions in Designing', *Design Studies* 13(2): 135–156.
Schurz, Gerhard, and Karel Lambert (1994), 'Outline of a Theory of Scientific Understanding', *Synthese* 101: 65–120.
Schuurman, Egbert (1972), *Techniek En Toekomst: Confrontatie Met Wijsgerige Beschouwingen*. Assen: Van Gorcum.
– (1990), *Filosofie Van De Technische Wetenschappen*. Leiden: Nijhoff.
Schwartz, Stephen P. (1978), 'Putnam on Artifacts', *The Philosophical Review* 87(4): 566–574.
– (1980), 'Natural Kinds and Nominal Kinds', *Mind* 89(354): 182–195.

Schweder, Rebecca (2004), *A Unificationist Theory of Scientific Explanation*. PhD. Department of Philosophy, Lund: Lund University.
Simon, Herbert A. (1973), 'The Structure of Ill-Structured Problems', *Artificial Intelligence* 4: 181–202.
– (1996), *The Sciences of the Artificial*. 3rd ed. ed. Cambridge, MA: MIT Press. Original edition, 1969.
Sober, Elliott (1983), 'Equilibrium Explanation', *Philosophical Studies* 43: 201–210.
– (1999), 'Putting the Function Back into Functionalism', in William G. Lycan (ed.), *Mind and Cognition: An Anthology*, Oxford: Blackwell, 63–70.
Stone, Robert B., and Kristin L. Wood (2000), 'Development of a Functional Basis for Design', *Journal of Mechanical Design* 122: 359–370.
Strevens, Michael (2004a), 'The Causal and Unification Approaches to Explanation Unified – Causally', *Noûs* 38(1): 154–176.
– (2004b), *The Sense of Understanding*. Book manuscript.
Tabery, James G. (2004), 'Synthesizing Activities and Interactions in the Concept of a Mechanism', *Philosophy of Science* 71(1): 1–15.
Thomasson, Amie L. (2003), 'Realism and Human Kinds', *Philosophy and Phenomenological Research* 67(3): 580–609.
Ullman, David G. (1992), *The Mechanical Design Process*. New York: McGraw-Hill.
Van Fraassen, Bas C. (1980), *The Scientific Image*. Oxford: Clarendon Press.
Van Inwagen, Peter (1990), *Material Beings*. Ithaca, London: Cornell University Press.
Verbeek, Peter-Paul (2005), *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Translated by Robert P. Crease. Penn State: Penn State University Press.
Vermaas, Pieter E. (2006), 'The Physical Connection: Engineering Function Ascriptions to Technical Artefacts and Their Components', *Studies in History and Philosophy of Science* 37(1): 62–75.
Vermaas, Pieter E., and Kees Dorst (2006), 'On John Gero's FBS-Model, Its Conceptual Framework and Descriptive and Prescriptive Aims in Design Methodology', *Design Studies* (forthcoming).
Vermaas, Pieter E., and Wybo N. Houkes (2003), 'Ascribing Functions to Technical Artefacts: A Challenge to Etiological Accounts of Functions', *British Journal for the Philosophy of Science* 54(2): 261–289.
– (2006), 'Technical Functions: A Drawbridge between the Intentional and Structural Natures of Technical Artefacts', *Studies in History and Philosophy of Science* 37(1): 5–18.
Vincenti, Walter G. (1990), *What Engineer Know and How They Know It*. Baltimore, MD: The Johns Hopkins University Press.
Volti, Rudi (1988), *Society and Technological Change*. New York: St. Martin's Press.

Walsh, Denis M. (1996), 'Fitness and Function', *British Journal for the Philosophy of Science* 47(4): 553–574.
Weber, Erik (1996), 'Explanation, Understanding, and Scientific Theories', *Erkenntnis* 44: 1–23.
Wiggins, David (2001), *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.
Wimsatt, William C. (1994), 'The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets', in Mohan Matthen and Robert X. Ware (eds.), *Biology and Society: Reflections on Methodology*, Calgary: The University of Calgary Press, 207–274.
Woodward, James F. (1997), 'Explanation, Invariance, and Intervention', *Philosophy of Science* 64: s26–s41.
– (2003), *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.
Wright, Larry (1973), 'Functions', *The Philosophical Review* 82(2): 139–168.
– (1976), *Teleological Explanations*. Berkeley, CA: University of California Press.
Ylikoski, Petri (2001), *Understanding Interests and Causal Explanation*. Ph.D. Thesis. Department of Moral and Social Philosophy, Helsinki: University of Helsinki.
Ziman, John, ed. (2000), *Technological Innovation as an Evolutionary Process*. Cambridge: Cambridge University Press.

# Ontwerpen reconstrueren, artefacten verklaren

### *Filosofische reflecties op het ontwerpen en verklaren van technische artefacten*

Dit boek gaat over *technische artefacten*: gebruiksvoorwerpen zoals paperclips, potloden, platenspelers en printers. Dingen die ingenieurs ontwerpen en die anderen vervolgens gebruiken om hun talloze praktische doelen mee te verwezenlijken. Wie geen bezwaar heeft tegen een wat mensvormige manier van praten over zulke objecten zou kunnen zeggen dat ze een verdeeld bestaan leiden. Aan de ene kant zijn het namelijk gewone fysische objecten, net zoals stenen en sterren. Aan de andere kant zijn ze onlosmakelijk verbonden met de intenties, doelen en plannen van mensen. Het zijn objecten die doelbewust ontworpen zijn voor een bepaald doel. Daarom schrijven we ook technische functies aan artefacten toe; we zeggen bijvoorbeeld dat lampen de functie hebben om ruimtes te verlichten. Verder spreken we normatieve oordelen uit over artefacten: 'Dit is een erg goed fototoestel.' Zulke uitspraken doen we niet over natuurlijke objecten.

Om artefacten volledig te beschrijven en begrijpen hebben we dus zowel een fysisch perspectief als een intentioneel perspectief nodig. Vanuit filosofisch oogpunt is het echter verre van duidelijk hoe deze perspectieven precies samenhangen – al sinds Descartes realiseren filosofen zich dat intentionele noties maar moeilijk in te passen zijn in een puur fysische beschrijving van de wereld en dat er dus een gespannen verhouding bestaat tussen het intentionele en het fysische perspectief op de wereld. De observatie dat artefacten desalniettemin thuishoren in deze twee perspectieven vormt het startpunt van dit onderzoek.

Om meer grip te krijgen op hoe beide perspectieven in artefacten samenkomen, concentreer ik me op de relatie tussen de fysische structuur en de technische functie van een artefact. Die relatie benader ik door studie te maken van twee contexten waarin structuur en functie op een systematische manier verbonden worden: het ontwerp en de verklaring van artefacten. Ontwerpen bestaat onder meer uit een overgang van een gewenste functie naar een concreet fysisch object (of een blauwdruk daarvoor). Als vervolgens verklaard wordt hoe een artefact zijn functie kan vervullen, wordt de fysische structuur van het

artefact gerelateerd aan de technische functie ervan. De centrale vragen zijn dus:

(1) Hoe wordt er in ontwerpen een overgang gemaakt van een beschrijving van een technische functie naar een beschrijving van een concreet fysisch object?
(2) Hoe worden technische functie en fysische structuur verbonden in verklaringen van hoe een artefact zijn functie kan vervullen?

Om hier meer over te zeggen is het nodig om eerst een aantal termen te verduidelijken. Dat gebeurt in hoofdstuk 2. *Ontwerpen* is een tweeledig proces waarin behalve een concreet object ook altijd een gebruiksplan voor dat object wordt ontworpen: een weloverwogen serie handelingen die tot de realisering van een praktisch doel leiden en waarin tenminste één handeling bestaat uit het gebruiken van het artefact. Gebruiksplanontwerp blijft in dit proefschrift grotendeels buiten beschouwing, ik concentreer me op dat deel van het ontwerpproces waarin een transformatie plaatsvindt van een gewenste technische functie naar een beschrijving van de fysische structuur van een artefact. Omdat *verklaringen* een veelbesproken onderwerp zijn in de wetenschapsfilosofie, kan ik niet volstaan met een korte karakterisering en wijd ik hoofdstuk 5 er geheel aan. De *fysische structuur* van een artefact bestaat simpelweg uit de materialen en componenten waar het artefact van gemaakt is en hun onderlinge relaties. De *technische functie* van een artefact vat ik op als het gedrag of de dispositie waarvoor een competente en welwillende ontwerper het artefact heeft gemaakt of een sociale gemeenschap van latere gebruikers het heeft geselecteerd. Met gebruikmaking van de zojuist geïntroduceerde notie van een gebruiksplan, kunnen we ook zeggen dat de technische functie de rol van het artefact in een gebruiksplan is. Artefacten kunnen in de loop van hun bestaan nieuwe functies verwerven en/of van functie wisselen, al zal dit zeker niet bij alle artefacten daadwerkelijk gebeuren. Daarnaast is het mogelijk dat een artefact een functie heeft maar die functie niet (meer) kan vervullen. Niets garandeert immers dat een artefact altijd kan doen waarvoor het bedoeld of uitgekozen is.

Met deze verhelderingen wordt het mogelijk om de vraagstelling verder te specificeren. De overgang tussen functie en structuur heeft drie verschillende aspecten.

(a) Het is een overgang van een beschrijving van gedrag naar een beschrijving van de componenten en materialen die dat gedrag mogelijk maken. Het is, met andere woorden, een overgang van een dispositionele naar een categorische beschrijving van een artefact. (Over

het onderscheid tussen dispositionele en categorische eigenschappen is in de filosofie enorm veel gediscussieerd. Voor hier volstaan twee punten. (i) Ik beschouw het onderscheid als een conceptueel onderscheid: dispositioneel en categorisch slaan op twee manieren om objecten en hun eigenschappen te beschrijven, niet op twee verschillende ontologische typen eigenschappen. (ii) Het onderscheid zit hem in het conceptuele verband tussen claims over disposities en claims over gedrag: wie zegt dat een object een dispositie heeft, zegt tevens dat dat object bepaald gedrag zou vertonen als het op een geschikte manier gestimuleerd zou worden. Wie een object beschrijft in categorische termen, committeert zich niet aan zulke claims over het gedrag van dat object.)
(b) Technische functies worden in eerste instantie vaak vrij abstract en incompleet beschreven met een korte karakterisering van het gewenste gedrag; een blauwdruk daarentegen is een zeer concrete en gedetailleerde beschrijving van de fysische structuur van een artefact.
(c) Technische functie horen thuis in het intentionele perspectief op de wereld. Ze vooronderstellen intentionaliteit omdat ze mede geconstitueerd worden door de intenties en overtuigingen van ontwerpers of gebruikers. Een beschrijving van de fysische structuur van een artefact is helemaal non-intentioneel en past prima in het fysische perspectief op de wereld.

Al met al is de overgang van functie naar structuur dus een overgang van een dispositionele, abstracte en intentionele beschrijving van gewenst gedrag naar een categorische, concrete en fysische beschrijving van structuur. Om die overgang goed te begrijpen moet dan ook aan al deze drie aspecten recht gedaan worden.

We kunnen al direct inzien dat in de context van ontwerpen het derde aspect van de overgang van minder belang is. De overgang van een intentionele beschrijving van gewenst gedrag van een artefact naar een niet-intentionele beschrijving van een concreet artefact is simpelweg een specifiek voorbeeld van de bekende overgang van iets willen doen of maken naar wat er uiteindelijk van terecht komt. Deze overgang is niet oninteressant – ze heeft alles te maken met de relatie tussen onze overtuigingen en de wereld, met (praktische) rationaliteit en onze cognitieve capaciteiten – maar het huidige probleem van het ontwerp van artefact voegt geen wezenlijk nieuwe elementen toe aan de problematiek.

Daarmee blijven aspecten (a) en (b) over: hoe komen ontwerpers van een abstracte beschrijving van gedrag naar een concrete categorische beschrijving van een artefact? Een logische plaats om te zoeken naar antwoorden op deze vraag lijkt het werk van ontwerpmethodologen. Ontwerpmethodologie is immers de discipline die zich bezighoudt met wat effectief ontwerpen is en

de middelen die je ervoor nodig hebt. In hoofdstuk 3 bestudeer ik daarom het werk van drie invloedrijke ontwerpmethodologen: Herbert Simon, Donald Schön en John Gero. Het resultaat hiervan is geen onverdeeld succes. Het blijkt dat deze ontwerpmethodologen maar weinig aandacht hebben voor de geschetste overgang. Wat ze er wel over zeggen is te vaag (Schön) of te generaliserend (Simon en Gero) om veel inzicht te bieden. Aan de andere kant blijkt wel dat ontwerpmethodologen veel praktische adviezen geven over methoden en technieken die ontwerpers dienen te gebruiken om hun werk effectief en efficiënt te doen. Als zulke adviezen breed gedragen worden door de gemeenschap van methodologen, ligt het voor de hand om te denken dat de aangeraden methoden en technieken daadwerkelijk bijdragen aan het maken van de overgang van functie naar structuur. Nadere studie ervan moet dan kunnen onthullen hoe zo'n methode of techniek dat doet.

In hoofdstuk 4 ga ik daarom dieper in op één zo'n techniek: functionele decompositie. Deze techniek beoogt het totale gewenste gedrag van een artefact op te breken in 'kleinere' of simpelere sub-gedragingen. Een rondgang langs verschillende auteurs laat zien dat functionele decompositie verschillende dingen bewerkstelligt: ze specificeert het totale gedrag meer in detail, legt een zekere mate categorische informatie over het artefact vast en maakt de beschrijving van het gedrag van het artefact concreter. Door het onderscheid tussen dispositionele en categorische beschrijvingen verder uit te werken en toe te passen kan een aantal van deze prestaties al beter begrepen worden, maar de kern van mijn analyse bestaat uit het onderscheiden van verschillende soorten relaties die er tussen gebeurtenissen (in dit geval gedragingen) kunnen bestaan. Deze verschillende relaties leveren verschillende vormen van functionele decompositie op. Functionele decompositie kan betekenen:

(1) Identificeren van de elementen van een verzameling gedragingen, waarbij die elementen samen genomen worden omdat ze bijdragen aan hetzelfde doel, op dezelfde of aangrenzende tijden of locaties plaatsvinden of simpelweg omdat ze uitgevoerd worden door hetzelfde artefact;
(2) Identificeren van de spatiële, temporele, of andere onderdelen van een gedrag;
(3) Identificeren van een superveniëntiebasis voor een gedrag.

Deze verschillende vormen van functionele decompositie dragen elk op hun manier bij aan het maken van de overgang van functie naar structuur. Ze specificeren het gedrag, maken de beschrijving ervan veel concreter en bepalen al een deel van de categorische beschrijving van het artefact. Over dit laatste wil ik nog iets meer zeggen, vooral omdat dit aspect vaak over het hoofd wordt

gezien in de ontwerpmethodologie literatuur. Ten eerste is het zo dat de sub-gedragingen die een functionele decompositie identificeert geordend zijn in ruimte en tijd. Daarmee is een klein beetje categorische informatie gegeven, namelijk de relatieve locaties van de componenten die de betreffende sub-gedragingen moeten gaan uitvoeren. Ten tweede wordt soms de input en output van sub-gedragingen behoorlijk gedetailleerd omschreven. Als dat gebeurt, zal een ontwerper met de relevante achtergrondkennis gemakkelijk kunnen afleiden wat voor soort component of onderdeel deze specifieke input-output combinatie heeft. Ten derde kunnen specifieke manieren om een gedraging op te delen in sub-gedragingen bepaalde (soorten) componenten of onderdelen suggereren. Iemand die ervaring heeft met deze componenten of onderdelen zal dat dan direct kunnen herkennen.

Met dit alles overbrugt functionele decompositie niet de gehele kloof tussen functie en structuur maar wel een belangrijk deel ervan. Als functionele decompositie klaar is, moet nog uitgezocht worden hoe de geïdentificeerde sub-gedragingen efficiënt en effectief gecombineerd kunnen worden en vervolgens geïmplementeerd in specifieke fysieke componenten en onderdelen. Hoofdstuk 4 geeft dus een partieel antwoord op de eerste onderzoeksvraag.

De context van verklaren vormt een spiegelbeeld van die van ontwerpen. Waar ontwerp draait om de overgang van functie naar structuur, gaat het bij verklaren grofweg om een overgang van structuur naar functie – de fysische structuur van een artefact moet althans een deel zijn van een verklaring van hoe een artefact zijn functie kan vervullen. In de hoofdstukken 5 tot en met 7 breng ik de verklaringsrelatie tussen functie en structuur in beeld en ook hier houd ik rekening met de drie aspecten van de functie-structuur overgang.

Hoofdstuk 5 bevat vooral voorbereidend werk. Ik geef daar een overzicht van de diverse theorieën van verklaren die wetenschapsfilosofen hebben voorgesteld. Het zou teveel afleiden om die hier te herhalen. Het belangrijkste is dat ik in hoofdstuk 6 gebruik maak van het causaal-mechanistische model van verklaren. Volgens dit model is het geven van een verklaring van een verschijnsel het beschrijven van een mechanisme dat dat verschijnsel (causaal) produceert.

Hoofdstuk 6 begint met enkele concrete voorbeelden van verklaringen die ontwerpers geven. Deze voorbeelden maken duidelijk dat zulke verklaringen laten zien hoe een artefact werkt. Het *explanandum* is de dispositie van een artefact om bepaald gedrag te vertonen, doorgaans het gedrag dat de technische functie vormt. Het *explanans* beschrijft de sub-gedragingen van het artefact en de fysieke componenten die die sub-gedragingen uitvoeren en laat zo zien hoe het artefact de betreffende dispositie heeft. Dit algemene idee werk ik uit

door twee verklaringsstrategieën te onderscheiden die *explanantia* genereren waarmee het gedrag *b* van een artefact *a* verklaard wordt: een *top-down* (TD) strategie en een *bottom-up* (BU) strategie.

> (TD) Beschrijf de sub-gedragingen, $b_1$, $b_2$, ..., $b_n$, van *b* en hun sub-gedragingen enzovoorts. Doe dit tot de sub-gedragingen op het diepste niveau voldoende basaal zijn. Geef ook de spatiotemporele configuratie van de sub-gedragingen. (Kortom: geef de functionele decompositie van *a*'s gedrag *b*.) Wijs tot slot aan welke componenten de basale sub-gedragingen uitvoeren.
> (BU) Beschrijf de fysieke componenten, $c_1$, $c_2$, ..., $c_n$, van *a*, de sub-componenten daarvan enzovoorts. Doe dit totdat je het niveau van basale fysieke componenten hebt bereikt. Beschrijf, voor zover mogelijk, de categorische bases van alle relevante disposities van de *ci*'s en verdere sub-componenten. Geef dan ook de functionele decompositie om te laten zien hoe zojuist verklaarde disposities samenwerken om het totale gedrag *b* van *a* te produceren.

Ik verdedig verder de stelling dat beide strategieën goede, op zichzelf staande verklaringen opleveren. Met name beargumenteer ik dat (TD) verklaringen niet incompleet zijn omdat ze minder uitvoerig zijn en niet alle categorische informatie bevatten die wel in (BU) verklaringen zit.

De oplettende lezer zal gemerkt hebben dat deze verklaringen weliswaar een verband leggen tussen een abstracte dispositionele beschrijving van een artefact en een concrete categorische beschrijving, maar nog niets doen met het derde aspect van de functie-structuur relatie, te weten het contrast tussen een intentionele en een niet-intentionele fysische beschrijving. Hoofdstuk 7 adresseert dit punt. De kern van dat hoofdstuk is het idee dat de verklaringen uit hoofdstuk 6 gecombineerd moeten worden met een functietheorie die de voorwaarden expliciteert waaronder artefacten een technische functie hebben, of waaronder functietoeschrijvingen gerechtvaardigd of correct zijn. In die voorwaarden zullen de doelen, intenties, plannen, overtuigingen en handelingen van ontwerpers of latere gebruikers een belangrijke rol spelen omdat, zoals ik aan het begin al zei, technische functies in feite geïntendeerde gedragingen van artefacten zijn. Verklaringen van hoe een artefact zijn functie kan vervullen kunnen dan de intentionaliteit en normativiteit van technische functies 'erven'. De verklaringen worden dan in het intentionele perspectief getrokken en het worden verklaringen van geïntendeerd gedrag en geïntendeerde disposities die geformuleerd zijn in termen van de intenties en/of overtuigingen die de ontwerpers of gebruikers ten aanzien van het artefact hebben of hadden.

Dat gaat als volgt in zijn werk. Wanneer we willen verklaren hoe een artefact zijn technische functie kan vervullen, moeten we twee gevallen onderscheiden. (1) De intenties van de ontwerper(s) of gebruikers zijn succesvol ten uitvoer gebracht en het artefact doet waarvoor het bedoeld is. In dat geval geeft de intentionele versie van de verklaring een correcte beschrijving van de werkelijke fysische structuur en disposities van het artefact. (2) Het artefact functioneert niet naar behoren, bijvoorbeeld omdat de intenties van de ontwerper niet succesvol uitgevoerd zijn, omdat het artefact kapot is gegaan, vanwege normale slijtage, of om een andere reden. In dit geval moet de intentionele versie van de verklaring – die, zoals gezegd, geformuleerd is in termen van de intenties en overtuigingen van de ontwerper of gebruikers – tenminste één onjuistheid bevatten. De ontwerper *dacht* bijvoorbeeld dat een bepaalde component ondanks trillingen op zijn plaats zou blijven zitten, maar dat is niet gebeurd. De intentionele versie van de verklaring geeft dus een (deels) onjuiste beschrijving van de werkelijke fysische structuur en disposities van het artefact in kwestie. Desalniettemin geeft zo'n verklaring nog steeds een correcte beschrijving van wat de ontwerper wilde of dacht dat de fysische structuur en disposities van het artefact zouden zijn.

In beide gevallen is het zo dat deze intentionele versie van de verklaring normatieve claims over hoe het artefact hoort te werken ondersteunt. Dat is vanwege de sociaal-epistemische arbeidsverdeling die er is tussen (professionele) ontwerpers en gebruikers van artefacten. Ontwerpers zijn 'artefact-experts' en het is hun verantwoordelijkheid om ervoor te zorgen dat de artefacten die ze maken correct kunnen functioneren. Gebruikers hebben vervolgens een sociaal-epistemisch recht (en in onze maatschappij tevens een juridisch recht) om de claims van ontwerpers over wat artefacten horen te doen en kunnen doen te vertrouwen.

Met deze combinatie van een functietheorie en mijn artefactverklaringen is de volledige relatie tussen functie en structuur in de context van verklaren in beeld gebracht. Daarmee is tevens de tweede onderzoeksvraag beantwoord.

# Curriculum Vitae

Jeroen de Ridder was born on April 4th, 1978 in Leerdam. He is married to Nelleke Sneep and together they have a daughter, Joa.

From 1990 to 1996 Jeroen received his VWO education at the Groene Hart Lyceum in Aphen aan den Rijn. He graduated *cum laude*. From 1996 to 2000 he attended Delft University of Technology. He obtained his M.Sc. (*cum laude*) in Systems Engineering, Policy Analysis, and Management there in 2000 with an exploratory study about the technical and socio-economic feasibility of quality of service mechanisms in the Internet. He was the best graduating student of the academic year 2000–2001 in the Faculty of Technology, Policy, and Management (TPM). His thesis was subsequently rewarded the NGI Informatics Thesis Prize 2000. Jeroen then attended the Free University Amsterdam and passed his propaedeutic exam (*cum laude*) in Philosophy there in 2001. (He is about to obtain an M.A. in Philosophy from the same institution at the time of writing this, which, he will readily admit, is a bit remarkable for someone who is also about to receive a Ph.D. in Philosophy.)

Before he started the Ph.D. project of which this thesis is the result, he held part-time positions as lecturer in the Policy Analysis Section of the TPM Faculty (2000–2002) and junior researcher in the Department of Philosophy of Erasmus University Rotterdam (2000–2001).

The work for this thesis was carried out in the Philosophy Section of the TPM Faculty at Delft University of Technology during the years 2001 through 2006. In this period, Jeroen spent three months as a visiting researcher in the Philosophy Department of Stanford University (2004) and another three months as a visiting student in the Department for History and Philosophy of Science of the University of Cambridge (2006). He also co-taught several introductory courses in the philosophy of science at the Free University Amsterdam in the same period.

He currently has a position as lecturer and researcher at the Free University Amsterdam.

## Simon Stevin (1548–1620), 'Wonder en is gheen Wonder'

This series in the philosophy of technology is named after the Dutch/Flemish natural philosopher, scientist, and engineer Simon Stevin. He was an extraordinary versatile person. He published, among other things, on arithmetic, accounting, geometry, mechanics, hydrostatics, astronomy, theory of measurement, civil engineering, the theory of music, and civil citizenship. He wrote the very first treatise on logic in Dutch, which he considered to be a superior language for scientific purposes. The relation between theory and practice is a main topic in his work. In addition to his theoretical publications, he held a large number of patents and was actively involved as an engineer in the building of windmills, harbours, and fortifications for the Dutch prince Maurits. He is famous for having constructed large sailing carriages.

Little is known about his personal life. He was probably born in 1548 in Bruges (Flanders) and went to Leiden in 1581, where he took up his studies at the university two years later. His work was published between 1581 and 1617. He was an early defender of the Copernican worldview, which did not make him popular in religious circles. He died in 1620, but the exact date and the place of his burial are unknown. Philosophically he was a pragmatic rationalist for whom every phenomenon, however mysterious, ultimately had a scientific explanation. Hence his dictum 'Wonder is no Wonder,' which he used on the cover of several of his own books.



## Simon Stevin Series in the Philosophy of Technology

*Books and Dissertations*

Volume 1: Marcel Scheele, *The Proper Use of Artefacts: A Philosophical Theory of the Social Constitution of Artefact Functions*

Volume 2: Anke van Gorp, *Ethical Issues in Engineering Design: Safety and Sustainability*

Volume 3: Vincent Wiegel, *SophoLab. Experimental Computational Philosophy*

Volume 4: Jeroen de Ridder, *Reconstructing Design, Explaining Artifacts: Philosophical Reflections on the Design and Explanation of Technical Artifacts*

*Research Documents*

Peter Kroes and Anthonie Meijers (eds.), *Philosophy of Technical Artifacts*

Technical artifacts are both plain physical objects and objects that have been purposefully made for a purpose; they have a physical structure and a technical function. As a result, they belong equally in a purely physical conceptualization of the world, in which human intentions and goals seem to have no place, and in an intentional conceptualization, which is used to describe and understand people and their mental lives.

This book explores how this observation plays out in the contexts of artifact design and explanation of how artifacts fulfill their function. It addresses the following questions: How do designing engineers get from a functional description of desired behavior to the concrete object that is the result of a design process? What do explanations of how an artifact fulfills its function look like and do they differ from explanations of natural systems?

*blurb: Lexicon Nr. 1 Roman A 9pt, white (line space 12pt); 2 columns, gutter 4 mm*

*margins and banner same as front side*

*'Wonder en is gheen wonder'*

*series motto: Thesis Sans Italic 12pt, 10 mm from top of banner*

*'Miracle is no miracle', Simon says*

TU Delft

TU/e technische universiteit eindhoven

*University logos (each 25% of column width), white 5 mm and 25 mm from top of banner*

9 789090 219035