AI that Glitters is not Gold

Requirements for meaningful control of AI systems

Kuilman, S.K.

DOI
10.4233/uuid:c2424141-fd13-473d-8511-4d5c6b4c492c

Publication date
2025

Document Version
Final published version

Citation (APA)
Kuilman, S. K. (2025). *AI that Glitters is not Gold: Requirements for meaningful control of AI systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:c2424141-fd13-473d-8511-4d5c6b4c492c

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

requirements for meaningful control of AI systems

SIETZE
KAI
KUILMAN

AI THAT GLITTERS IS NOT GOLD

AI THAT GLITTERS IS NOT GOLD

REQUIREMENTS FOR MEANINGFUL CONTROL OF AI SYSTEMS

AI THAT GLITTERS IS NOT GOLD

REQUIREMENTS FOR MEANINGFUL CONTROL OF AI SYSTEMS

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op Woensdag 12 November 2025 om 10:00 uur

door

Sietze Kai KUILMAN

Master of Science in Artificial Intelligence, Universiteit van Amsterdam, Nederland
Master of Arts in Philosophy, Universiteit van Amsterdam, Nederland,
geboren te Vlaardingen, Nederland.

Dit proefschrift is goedgekeurd door de

promotor: prof. dr. C. M. Jonker
copromotor: Dr. L. C. Siebert
copromotor: Dr. S. Buijsman

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. C.M. Jonker,	Technische Universiteit Delft
Dr. L.C. Siebert,	Technische Universiteit Delft
Dr. S.N.R. Buijsman,	Technische Universiteit Delft

Onafhankelijke leden:

Prof. dr. J. vd Hoven	Technische Universiteit Delft
Prof.dr. H.B. Verheij	Rijksuniversiteit Groningen
Prof. dr. B. Haring	Universiteit Leiden
Dr. G. Mecacci	Radboud Universiteit Nijmegen

Keywords: Entrenchment
Meaningful Human Control
Self-reinforcement
Relevancy

Printed by: proefschriftspecialist.nl

Front & Back: Dirk V.d. Vijgh

SIKS Dissertation Series No. 2025-55
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN 978-94-93483-21-7

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

In memory of Bert

CONTENTS

SUMMARY

Under which conditions can you say that a system is actually and meaningfully under your control? Accidents happen with machines and often that is not the fault of the user. So what does control entail? To gain some modicum of understanding, we need to learn how technology and control relate to one another. Of course, the concept of control also relates to the consequences of control, namely the responsibility you have for that which you can control.

On average, we are used to thinking about technology as a kind of hammer, something on hand that we can make use of. Yet, such a hammer also invites us to hammer things. This invitation is a kind of mediation through which we are encouraged to act in one way and not another. In short, technology can also influence us.

If technology can exert a kind of influence on you, then we need to ask what that means for control. What kind of issues do we run into because of that influence? In this dissertation I investigate a few key issues: self-reinforcement and relevancy.

Technology tends to entrench itself in society once it is widely implemented. This process of entrenchment is often through self-reinforcement (chapter 2). Like a snowball rolling down a hill that picks up more snow as it goes, so too can technology gain a kind of traction that becomes harder and harder to ignore and disband or even change. Consider, for example, how much has to change if we want to live without cars. The moment technology gets picked up at large, we also institute policies and create institutions around which such technology can be legitimized. The point is that the technology can create a new standard to which everyone grows accustomed.

The problem of relevancy (chapter 3), however, moves beyond the domain of application and more into the domain of design. How is a designer supposed to know all the relevant details to incorporate all the necessary stakeholders, values, and knowledge into a design? A highly improbable feat, unless one believes the universe to be calculable and we are given infinite time.

A side effect of entrenchment and relevancy together is that we may find it harder and harder to alter technology even if it systematically disadvantages particular groups. Technology promises to improve our lives, but such a promise is not made in a vacuum. If we take entrenchment and relevancy together, if they are embedded and difficult to change and one is ill-suited or ill-adapted to the technology, it basically boils down to an additional barrier (chapter 4). If this is compounded by an inability to contest the technology, it may also be embedded in how we view the world around us. E.g. we start to

accept cars as a fact rather than something about which we can debate.

Given that we have these kinds of costs, what can we do to curtail them and exert more control over technology? Well, one way to look at control and technology is through the lens of Meaningful Human Control, which can result in a tracking and tracing condition. Tracking means keeping track of all the relevant reasons, such that the system knows how to act appropriately in a given situation. Tracing means we need to be able to trace towards a responsible person if something goes wrong, given the situation and choices made. Both of these seem difficult to do if technology can influence you in the way I just described. Are there ways forward?

I provide means of improving and concretizing these conditions. On the one hand, we could look more at limiting these technologies towards a particular user base (e.g. limiting the size of users allowed on the platform) (chapter 2). This way you can dampen the problem of entrenchment. You could place more emphasis on the responsibility of the designer, and you could try to incorporate contestability into the lifecycle of the system (chapter 3). That incorporation of contestability has to be in such a way that we can incorporate different reasons and ideas about what is relevant, such that the system gets meaningfully adapted.

Yet, even the application of Meaningful Human Control in design is not easy. While the tracking conditions rely on relevancy (which is not a given), the tracing condition tends to rely on causal and epistemic conditions, which seem ill-fitted to give an accurate account of responsibility (chapter 5). To correctly apply tracing conditions, we fundamentally need normative conditions to trace towards a responsible individual. As it appears that both epistemic and more performative conditions surrounding responsibility are problematic when applied to the design domain because they cannot accurately account for unknown unknowns. Leading to questions of whether a designer ought to be accountable for the actions of a user, who was invited to act in a particular way but not forced.

The application of normative conditions throughout design, implementation, and policy may aid in the actual improvement of control over technological systems and curtail some of the costs that coincide with them.

SAMENVATTING

Wanneer heb je controle over een apparaat? Onder welke voorwaarden kun je zeggen dat een systeem dat ook werkelijk onder jouw toezicht staat en naar je luistert? Als er ongevallen gebeuren door machines, is dat veelal niet te wijten aan de fout van een gebruiker. Wat moeten we daarmee?

Gemiddeld genomen zijn we geneigd om over technologie na te denken als een soort hamer, iets wat wij in de hand hebben en waar we gebruik van maken. Echter, die hamer geeft ons ook de kans om dingen te zien die we kunnen hameren. Er zit dus ook iets in dat ons beïnvloedt. We worden uitgenodigd om op een bepaalde manier te handelen en daarmee beïnvloedt het ons.

Als het ons beïnvloedt, dan moeten we ons afvragen wat dat betekent voor controle. Wat voor problemen levert het op als we worden beïnvloed? In dit proefschrift onderzoek ik een aantal van deze zaken: zelfversterkende effecten en relevantie.

Het is vrij zichtbaar dat technologie zich een weg weet te banen door de maatschappij en zichzelf overal nestelt wanneer het wijdverspreid gebruikt wordt. Hoe deze vorm van innesteling gebeurt, komt mede door zelfversterkende effecten. Zoals een sneeuwbal die van een heuvel afrolt en daarom meer sneeuw meeneemt, zo kan technologie ook grip krijgen op de maatschappij en zo vastgeroest raken dat het moeilijk is om over alternatieven na te denken. Denk bijvoorbeeld aan hoe een wereld eruit zou zien zonder auto's. Dat is te waanzinnig om te bedenken en toch is het brede gebruik van auto's redelijk nieuw. Het moment dat machines in beeld komen, hebben we ook vaak instituties en beleid nodig om het in goede banen te leiden. Tegelijkertijd legitimeren instituties en beleid de machines. Het punt is dat technologie een nieuwe standaard maakt waaraan iedereen gewend raakt, of het nu in je voordeel is of niet.

Het probleem van relevantie gaat voorbij aan de toepassing en meer over het ontwerp. Hoe dient een ontwerper alle relevante details in acht te nemen? Zijn alle verschillende partijen en perspectieven wel aan bod gekomen? Om dit in één keer goed te doen, lijkt onmogelijk. Tenzij je gelooft dat alles berekenbaar is. Dus als dat niet in één keer goed kan gaan, hoe kunnen we dat dan beter doen?

Een bijeffect van innesteling en relevantie is dat ze het moeilijker maken om technologie aan te passen, zelfs als het systematisch een groep benadeelt. Technologie heeft als belofte dat het onze levens bevordert, maar zo'n belofte gebeurt niet in een vacuüm. Indien we innesteling en relevantie samenne-

men, dan verschijnt het probleem dat technologie voor sommigen niets anders is dan slechts een volgende barrière. Als we daarbij niet de kans krijgen om ons te verzetten tegen technologie, dan leren we het misschien accepteren als een feitelijkheid, terwijl dat eigenlijk niet zou moeten.

Gegeven zulke problemen, kunnen we wat doen om de zaken te verbeteren? Een manier waarop we hiernaar zouden kunnen kijken, is door het gebruik van *meaningful human control*, waarbij we specifiek kijken naar de vertaling tot *tracking* en *tracing* condities. Tracking (volgen) betekent dat we de relevante redenen in acht nemen, zodanig dat het systeem gepast handelt gegeven een situatie. Tracing (traceerbaarheid) betekent dat we handelingen kunnen herleiden tot een verantwoordelijk persoon. Beide condities zijn echter moeilijk toe te passen als we de problemen van net in acht nemen.

Ik geef een aantal manieren waarop we zowel de condities kunnen verbeteren als kunnen concretiseren. Enerzijds kunnen we denken aan het verkleinen van de invloed van technologie. Denk bijvoorbeeld aan een maximale hoeveelheid gebruikers. Zo kunnen we innesteling verminderen. We kunnen ook meer verantwoordelijkheid leggen bij ontwerpers. Ook kunnen we weerstand beter proberen mee te nemen in de cyclus van AI-ontwerp, zodanig dat we beter kunnen luisteren naar andere relevante redenen.

Echter, zelfs met meaningful human control hebben we nog steeds problemen. Alhoewel zulke condities afhankelijk zijn van relevantie (wat al een probleem is), lijkt de huidige versie van traceerbaarheid te onzorgvuldig vormgegeven. Als we hiernaar kijken, komen we uit op een probleem waarin causale, epistemische en zelfs performatieve condities tekortschieten om een stevige grip te geven op wie nu eigenlijk verantwoordelijk is. Om daarmee om te gaan, zullen we normatieve omstandigheden moeten toevoegen. Hierdoor lijkt het erop dat de invloed die technologie heeft op gebruikers meer de verantwoordelijkheid is van hen die van die invloed winst halen dan van iemand anders.

ACKNOWLEDGEMENTS

One of my heroes, Raymond Carver, wrote and said a few things about writing in his life. One of the most important details was: *get in, get out, don't linger*. Kurt Vonnegut, another hero of mine, said something along a similar vein: *your reader's time is valuable*. While Carver talks about keeping things brief because he had little headspace for a long novel, Vonnegut, instead, considered all the distractions around while readers took their time to read what you had written down. So he thought that your writing had better be worth it. I have a habit of overdoing both Carver and Vonnegut. I take far too little time, preferably getting to the conclusion of my ideas before all the words have hit the page. Impractical and unreadable, is the concise way to put it. To that end, Carver also taught me that writing is more than expression, it is also the act of *communicating to your reader*. And that act of communication I have always found much harder, mostly because I think in leaps and bounds, crossing roads that I see as logical, but that tend to be incomprehensible to others. So, without the serious help of lovely individuals, I would not have been able to write anything truly tangible, and I would have never been where I am today.

Luciano Cavalcante Siebert, thank you for giving me the opportunity to write this thesis, for sparring, commenting, and debating with me the countless ideas that I posited throughout the past four years. I got to know you as a very kind person, who has always given sufficient leeway and always made sure I took care of myself first. I can't imagine any other daily supervisor that would have been a better fit.

Catholijn Jonker, I admit I have enjoyed the freedom and honesty that you've provided over the years, both in criticism and in openness about the practice of academia. I have truly found it a privilege that my radical and theoretical disputes were not met with dogmatic appeals, but rather led to fruitful questioning.

Stefan Buijsman, thanks for reigning in my more continental tendencies and for joining the project. The interaction with you was often highly efficient yet playful. You've given structure to my thoughts and taught me how to be more explicit. I will enjoy those things for years to come.

All three of you have given me a time at Delft that I have not only cherished as a researcher but also fundamentally as a person. I couldn't have wished for a better supervisory team.

I have also been surrounded by a few loving friends who invested the time into listening and understanding what I was trying to do: Juun Ly, Thomas Van

Zwol, and Robin Riemersma. Between the three of you, each idea within this booklet is as much yours as it is mine. Robin, your thoughts on philosophy and the countless discussions we have had on a multitude of topics sent through philosophy in light speed. There were always more books or ideas you suggested, and it has given me a myriad of treasure troves to plunder. Thomas, our discussions on artificial intelligence and your openness to my alternatives and criticisms, has given me much insight into communicating my thoughts to other computer scientists. Juun, you have seen more of my issues with writing a thesis and the obstacles of academic life than anyone else. The mix between both intellectual understanding and introduction to topics I would otherwise not have considered (looking at you, Wicked Problems) has been a blessing. I know I would never have committed to this project well without any of you.

Furthermore, I need to thank Sven Nyholm. Thank you for hosting me so generously at the LMU in Munich and, in general, for taking the time to listen to me early on in my PhD as well. I always went home feeling heard in a way that I truly appreciate. You never dismissed any of my questions and were willing to honestly answer what an academic life looked like. In that vein, I would also like to thank many of the others in Munich who made me feel welcome.

I am grateful to Roel Dobbe, Jeroen van den Hoven, Filippo Santoni de Sio, and Herman Veluwenkamp. I loved the discussions, whether it was on content or simply on what academia was like.

Thanks to everyone from the DDEC, you've always made me feel very much at home. I enjoyed the dinners and writing retreats we've had.

Zuzanna Osika, thanks for jabbing through the facade and always being there to talk to in the office. You've become a good friend. Michaël Grauwde, thanks for the good reads and the fun conversations. Mohammed Al Owayyed thanks for the conversations about games and the suggestions. Antonio Mone and Amir Homayouni Rad, thanks for always being in the office. I hope both of you will find your way in the world of commerce and keep telling me about it. Many others have had an impact on this dissertation. Either through small discussions or simply by aiding me in some form or another. My gratitude goes out to: Nora Angelys, Naud van Beuningen, Wendy Boeree, Mia Brandtner, Bas Broers, Hans Cronau, Rens Cronau, Marianne Franken, Jeroen Frans, Stefan Gaillard, Lara Hakkenes, Nick Johnston, Amber Knoth, Lorijn Lammes, Marlou van Manen, Reinoud van Meerwijk, Maartje Meesters, Siddharth Mehrotra, Thomas Meier, Auke Montessori, David Nayer, Jasper Nijssen, Roel Nijssen, Kyrke Otto, Nanda Pellikaan, Miny Raijv, Evelien Renders, Maarten Rood, Anne van Rhijn, Nynke van Uffelen, David Verhoef, Florens Vernooij, Ralitsa Videnova, Benny van der Vijgh, Dirk van der Vijgh, Muriel van Winckel, Emma van Zoelen, as well as the staff of II and many others.

I also don't think I would have got through this PhD without sport, though

the concussion really wasn't helpful. Yet, I do want to thank all the great people from Allroundgym for the good times, as well as the people from Pallos.

I am also grateful to my parents, Karel and Karin, who set me on this path within academia, and to those who I count as siblings: Wilmar, Fedde, Jorine, and Arend, who shaped me in many more ways than I can imagine.

And to some (you know who you are): Thank you for sticking by me through sickness and through health, and for helping me remain steadfast in my convictions. I love you very much. There are no words to express your importance to me.

PREFACE

With this dissertation, I have attempted to do something that belongs to neither philosophy nor computer science. As a student of both fields, I had truly become dismayed with the range and preached potential of artificial intelligence and its purported scientific methods. I had aims and ambitions to clarify the problems that originated from the way engineers developed their ideas of science, specifically taking old philosophical debates and showing their current-day implications. As I went through this PhD, however, I learned that philosophy too has its shortcomings regarding computer science, as its scientific discussions are often spent in ivory towers, far away from reality. While philosophers do engage with some parts of computer science, they tend to bring their roots along with them and see only the ideal solution to a problem rather than the real-world implementation. Even if such discussions are nourishing, they may not appeal to an engineering audience. This means that a wealth of information is lost on both sides.

My quest has been to show how we can bridge that gap in a meaningful way, and that makes this a cautionary tale. What I learned while trying to bridge that gap is that transdisciplinary work is hard. One is easily caught up and swept away by the current, meaning you start building something on either side of the gap without ever intending to close it. In trying to do so, I started to see why leaping from one field to the other is much easier. Engineers build, while philosophers ponder. An engineer works through an iterative process of improvement and knows the start is wrong but has the potential to change it later. A philosopher works through a process of idealizing and refinement; getting a good definition down and good questions in the front of one's mind is essential. To go at it willy-nilly as a philosopher is most likely going to end up reinventing the wheel and a coinciding discussion that happened a hundred years ago, if not much longer ago. The two are more incompatible than one may first expect. To ask an engineer to lay down its tools of the trade and think about disagreement, perspective, and problem formulation, fundamental issues about the truth of a certain metric and its genealogy is bound to cause issues. A philosopher who codes is bound to make something needlessly complex because of the potential edge cases and variety of views and beliefs one may have about the underlying assumptions of a system, or even worse, creates a terrible implementation because of certain abstractions made early on.

To give an idea about how to think about philosophy and computer science, we can start with a simple example taken from real life. A good friend of

mine works at a bank as a data engineer. They process heaps of data every day and are more like a middleman than anything else. If they screw up the ordering of the data, then everybody down the line has issues with it. Thus, they are cautious. They write tests first in their development environment and try to create all the necessary tests such that they have the important bases covered. After that, they go about iterating and implementing and seeing whether their implementation fails their defined tests. It is a simple premise, but philosophy, I do believe, can provide such meaningful requirements, surface potential discussions, and go even beyond it. Philosophy has the potential to go beyond mere implementation problems and show, inherently, a way to change one's mind and perspective on the things perceived. On that level, it has the capacity not only to help build safer technology, but even to provide fundamentally alternative routes and futures. One such simple example is one of the problems discussed in this dissertation: The Frame Problem. In essence, it explains why first-shot solutions are likely to run into problems and why simulations are either overly complicated or overly simple. A philosophical problem can help us understand that technology needs to be tested and arranged in a particular way before even putting it out in the wild, and even then, we need to be aware of our potential limits of simulation.

Yet, computer science is a merit to philosophy too. The introduction of new technology is not merely a showcase of our technical prowess. It can help us think critically about what kind of society we want. The change in society, the change to a life, and the introduction of capacities never seen before can give food for thought in terms of what humanity, the good life, and society should look like. And in fairness, it allows philosophers to earn their bread and squabble over the limits of consciousness and the mind. During my PhD, we all became acquainted with ChatGPT, which introduced new questions for me and for many other teachers as well. Students asked whether it was alright to use it, and that seemed like a moral gray zone. I simply thought of it as lazy, as a way to cheat yourself out of learning some necessary skill in life (e.g. writing), but nonetheless we were confronted with a change in society, and a debate about what would be the right use and the right move forward.

Working in between computer science and philosophy, made me aware that it always felt easier to belong to either one group rather than both. I may have been taught to think like a philosopher, yet when one is busy working out the bugs in the code, there is simply little head-space to consider many of the larger implications and theoretical limits of the systems one is building. Instead, it always felt to me as if those limits suddenly hit me. Blinded by oversight, I paralysed during the process. It felt as if two entirely different worlds could collide within me. What I could assess during my PhD was the fact that truly transdisciplinary work is like trying to teach a language that doesn't exist. You are constantly translating for everyone in the hopes that they can grasp the meaning of the propositions you represent. But translation is of course

always adaptation, and with adaptation we may lose some ideas and nuance essential to good understanding.

It dawned on me that the reason for this language barrier is rather simple. Philosophy is a deep well of understanding and discussion, disagreement and tradition. Artificial Intelligence is the young hot hype, that has only recently entered the halls of the academia. The people who are taught about artificial intelligence, are taught little if at all about the automatons of yesteryear, the automatic writing machines, Babbage, or Lady Lovelace. We aren't even teaching that much logic any more. ChatGPT came onto the scene only two years ago, and it has dramatically changed how we view these kinds of technology as well. Historical context and continuity gives discussion a particular shape, a particular movement and speed. The scientific field of artificial intelligence does not have a great track record when it comes to slow reflection. In artificial intelligence, everything has to be recent. New benchmarks show the continual progress of performance. It's a process of keeping up with the digital Joneses. The field has a tendency to look forward. It is determined by progress and improvement, preferably with a never-ending acceleration. The caution and slow thinking of philosophy, with its traditions, is a hard bargain for those who think that a discussion from the eighties is ancient. Let alone the worth, one may still uncover in the works of Aristotle, Schelling, Wittgenstein, and many others. Yet, there are things we should speak of as a community of computer scientists and philosophers. The political power engendered by such systems, the concepts of meaning as generated by machinery, the equivocations and anthropomorphism that comes so natural to the way we speak of machines. Nonetheless, the way we are taught makes caution and slow thinking simply incomputable. We are building to see if certain things are possible, and not whether it is preferable to do so.

It is exactly those questions: *what is possible? And what is preferable?* at which both engineers and philosophers can meet, but also where the difficulties lie. Before I ended up in Delft, I wrote two theses about the impossibility of algorithms grasping any mode of understanding. In part, I drew inspiration from a discussion about what can be meaningfully said. It interlinks a kind of understanding, with language and its potential limits within the world. Many philosophers have grappled with similar questions throughout history when it comes to the possibility of language. Frege comes to mind as he grapples with the inscrutability of thought without the lens of language, but I am also reminded of Hume's ruminations on human understanding, the hands of Moore, Russell's ideas about a French king, Davidson's triangulation, and Wittgenstein's silence. All of them ask a question about language which is seemingly far removed from computer science. Yet, we only need to reframe the way we think about "what can be said."Our potential to create artefacts (which I will take to mean the broadest category of objects under which computers and algorithms are merely a subset) are - as I believe - part of a similar discussion.

Artefacts, like language, can seem meaningful, and we get confused when we start to believe that sentences like: “Green dreams sleep furiously"have meaning, or when we believe we can use machines to predict the future. Artefacts, too, are spoken by an interlocutor who states them in a particular way and with a particular goal. They are like a password which can deny access to some. As such, artefacts can deny access to those who cannot or will not wield them, or whose features are not portrayed by such things. Worse is that such machines are seemingly objective, they provide a sense of certainty, or in the worst case, the promise of certain knowledge. The conflation of meaningless artefacts and meaningful ones, is difficult because we like to listen. They are deceptively well worded.

Nonetheless, possibility does not equate to preference. We only need to be reminded of doublespeak and the terror of language to see the mind-bending nature of it. Hannah Arendt provides a marvellous example of the saboteur in the Soviet communist party. The saboteur states that he knows or is sure that he is not one, but if the party says, so it must be true. A first-hand account of that came for me with the childcare benefit scandal[1] - in which a woman simply stated continually stated she was dumb and that it was her fault - she could not fathom that the system was perhaps wrong. While they are not entirely the same, artefacts can decide for us, structure millions of results in fractions of seconds, and we accept their truth sometimes in eerily similar ways as the truth of the Soviet party. Both have the greatest potential to deny our own experience of the world and to invite thoughtlessness. The artefact of doublespeak is one in which engineers and machines promise ground truth, or benefits, like connecting to everyone, while immediately enslaving us to addictive technology, which polarizes us, which separates us, and which harms our potential to view the world in new and unexpected ways. These things can make us careless and dependent in a world that becomes less human and human compatible.

The fact that language and artefacts distort and manipulate, can be viewed from three different angles. The first is about oppression through technology and its political implications, much in the same way overpasses can be racist. In design, this is a prominent example. As Robert Moses, a designer, purportedly created overpasses that were too low for the bus. Because mostly people of colour used the bus, they were denied access to particular parts of the city. In this discussion of politics we need to wonder, as Kobi Leins puts it: “whether AI is at all suitable, or compatible with democracy, for certain functions."This discussion is actively held by a lively community of philosophers, and they are becoming harder to ignore.

The second is the way we may alter our perception through language and

[1] The childcare benefit scandal was a large scandal in The Netherlands in 2020, in which parents were unjustly classified as frauds, this caused numerous problems and harm. It was in part caused by an algorithm which discriminated against race.

technology. Ideas of freedom can alter the way we view the world. It means we can perceive it differently. In philosophy of technology, this is mediation theory and technomoral change, which we will see later down the line. In understanding of texts, we see something like the hermeneutic circle, in which our understanding is altered by the texts we consume. Our horizon can subtly shift due to the introduction of new information. An easy way to conceive of this is to see how our behaviour changes once we grasp a new idea, or when we finally understand how to work with a new piece of technology. We can be constituted differently because of those things. For example, we need *freedom* if we want to conceive ourselves as *free*. So, too, do we need *computers* to think of ourselves as *computer scientists*.

The third way we can look at language is the way language can distort and manipulate in terms of its own limits. There are nonsensical things that can sound like they have meaning, and this, I believe, is the ultimate question to ask about technology: What can be meaningfully erected as an artefact? This, I propose, is also the underlying thought throughout this dissertation. To understand that question is also the work of a lifetime, not merely four years and a doctorate degree. I have shifted most of the discussion in this dissertation to a moral question, as that is far more palatable to engineers. Yet, I truly believe there is an element of technology itself which is about *meaning*. What are the limits we ought to adhere to make an artefact legible and legitimate? One central concept that came to pass while writing this dissertation was the concept of values, and the idea that computers ought to act according to ours. We can talk of political values, but the larger question that looms in the background is whether machines can portray our values at all, whether we are capable of formalizing these values such that they are portrayed in a meaningful fashion. Again, we can see a similarity between language and artefacts. We may have difficulty putting down our thoughts and feelings in words, there may be things that are completely undescribable otherwise me may have difficulty pinning reference to such an extent that it actually covers the ground we want it to. Why are we so quick to think that the world will be subjugated to our will through the use of such categorizations and formalizations in machinery?

If philosophers and computer scientists see eye to eye, it should be about these kinds of questions. Between preference and possibility lies a small window of opportunities that make the world a better place. In language, it sometimes seems that truth may elude us, as the link between sense and reference is difficult to discern. In much the same way, technology has a similar gap between design and implementation. We seem incapable of explaining the links between words and the world - that are not open to alternatives - and similarly, we seem incapable of designing algorithms that aren't open to use in unintended ways. People run away with things we said or twist our words. Thus, how can we meaningfully speak and erect artefacts? Well, they both seem to depend on bridging a gap, and it is a question of whether that can be done and if

not, by what means we can attempt to be better.

Sietze Kai Kuilman
Delft, August 2025

1

INTRODUCTION

We have got onto slippery ice where there is no friction and so in a certain sense the conditions are ideal, but also, just because of that, we are unable to walk. We want to walk so we need friction. Back to rough ground!

Ludwig Wittgenstein, *Philosophical investigations*, §107

1

1.1. CONTROL, USE, AND DESIGN

In 2018, a man was found guilty of eavesdropping on his estranged wife. It was a peculiar crime because he used technology to do it. Specifically, he used the microphone of a wall-mounted tablet. The tablet was not at all meant for that, rather it was used to control the heating and lights in their home. Nonetheless, it did have that capacity. Listening to her in secret this way eventually caused him to become aggressive [1].

When looking at such a situation, we tend to argue along a certain line of thought. Technology is not about *what it is*, but rather *how it is used*. Remote controlled devices, if left unattended and taken together with someone's malicious intent, can really make a house an unwelcoming place. Suddenly, someone can turn on the lights in the middle of the night, or to turn up the volume of a speaker, put on music, and so forth. Smart appliances in the home can be downright abusive. But if they are used properly, they can be an easy way to control one's home.

The question is whether such a narrative is correct. Of course, the moral of the story can be viewed in a myriad of ways. We can blame the man who used the appliance inhospitably. We can also shift the blame onto the tablet and its manufacturer for providing such a microphone at all, which would lead us to the lack of protection these companies proffer. Or we could posit in terms of a societal scope where the lack of governance on this fact should be blamed. After all, for some, it may be the task of the government and society to protect us from such harms. All of these positions say something about the technology, its goal, and its purported use. If we plant the blame on the manufacturers, then we could say the technology was obviously ill-designed, flawed even, because they gave way for such a terrible thing to happen. Could we have designed it differently? Often times, misuse causes us to go about improving the technology. Which suggests that design and use are related. For smart appliances in the home, we desire to become aware of these harmful potentials and see whether we can design them differently [2, 3]. However, how can we design complex technology differently?

While there are endless opportunities to look into the issue at hand, there is one essential aspect of technology bound up with use, and that is: control. For example, if I want to use a hammer, I need to be able to control it such that I don't end up hitting a person instead of a nail. I must be in control of the tool if my use of it is valid, meaningful, and effective. I would rather not hit colleagues on the head with a dangerous implement, nor do I want to misuse my power. Furthermore, I would like to make sure that a hammer isn't used against me. My control may be diametrically opposed to your control.

Now, many new technologies like AI systems are hazardous in this regard. The story we like to spin is that it is all outside of our control. This is the classical: *Computer says no!* from the TV show Little Britain. Yet, what we need to look at critically is whether that *out of our control* is justified or not. Did we

not design the system that made it say no in the first place? When is such an answer valid, and when is it not? The problem of current AI systems is debated quite heavily. We see that people are worried about mental health issues in relation to social media [4]. There are issues of engrained bias, racism, and sexism in machinery [5]. We worry about responsibility gaps [6, 7], about sex robots and their effect on sexuality and feminism [8, 9]. We know that they may harm democratic institutions, they are being called weapons of math destruction for a reason [10]. So when do we control these systems, and can we prevent the problems that follow from that? Furthermore, we must also ask whose weapons they are, whose values we embed, whose problems we are solving. In short, who are we putting in control? The question of control is fundamental to much of the debate in philosophy of technology. In The Social Control of Technology [11], Collingridge already argues that we are faced with a fundamental issue of control. Namely, we do not know the impact of technology once implemented, but at that stage it is so entrenched that it is difficult to adapt, remove, or change. But this entirely begs the question: Who is the invested party that wants to keep it entrenched?

This investment and entrenchment is a kind of normativity. The debate surrounding the entanglement of normativity and technology is found in a plethora of discussions. Kranzberg, for example, famously wrote: Technology is neither good nor bad; nor is it neutral [12], alluding to the fact that there is some kind of normativity involved. In mediation theory by Verbeek [13] we learn that our perception is altered by the introduction of technology, which pulls into question what it means to be *us*. The question of who we ought to be quickly follows. An example in philosophy of technology posits design as: what is desirable and technically feasible [14]. More obviously tied to artificial intelligence technologies, we see the question of value alignment [15]. With such technology specifically, we want them to act in accordance with our values. Mostly because of their capacities. Such a debate about value alignment has already led to a myriad of solutions. The improvements for technological development are bountiful. One good example is the debate about meaningful human control [16–20], in which we argue that we should ultimately have control over such systems and how that ought to be done. There are also concepts like human-in-the-loop, human-on-the-loop [21], in which the human is an active participant in some form or another and able to exert some version of influence. Otherwise, there is talk of value-sensitive design [22], in which we try to incorporate values during the design, as well as inverse reinforcement learning [23], in which we try to infer the correct policy from behaviour.

Given that there are already so many discussions about improving our technology and the normativity that goes with it, what is there to say? Well, the entire debate is far from solved. Feenberg [24] for example, distinguishes four different strands of thought about technological development: determinism, instrumentalism, substantivism, and critical theory. Each of these diffe-

rent categories suggests something about how value-laden and autonomous technology is. In deterministic and substantivist thought, we see that technology in and of itself is, to a certain extent, autonomous and therefore uncontrollable; the difference is that substantivists argue it is rather value-laden. By autonomous technology, I do not mean something like self-driving cars, but rather that we have no control over what new technology will be developed and why. The determinist will say, *it is going to come to society anyway, so we better adapt to the technology.* Instrumentalism and critical theory let us steer towards alternatives, positing that we are capable of controlling the flow of technology. Again, the difference between them is that instrumentalism argues technology is value free and critical theory does not make such an assumption.

Furthermore, a lot of the debates surrounding normativity go into detail about *improving* the technology, rather than asking whether it is appropriate to implement it at all [25]. Rather than assume we *ought* to improve technology, we can instead look at design requirements that belong in the technological development itself. In doing so, one stops to separate the design process from the system in which it is nestled, but we can start to genuinely ask, should it be here in the first place?

If we are to investigate technological development from the perspective that human beings are in control and that the technology is value-laden, then we need to understand two different concepts: control and wickedness.

1.2. THE BASICS OF CONTROL

If you are willing to look for it, you'll find that the notion of control has left its claw marks on literature and popular culture. These impressions can actually provide a fairly good roadmap to understanding control. What I am providing in this section is an intuitive sense of control and some basic concepts that will be helpful later down the line. What it means to be in control of someone and something, but even more pressing, what it means to be in a position without it. The reason for giving this intuition is that control is little discussed on its own [26]. Rather, it is discussed with coinciding concepts or factors. For example, there is talk of responsibility and control [27], power and control [28], or control over technology [11], control is also a part of action theory [26]. We are going to delve into the basics, as it is the most promising way of understanding the problem at hand (namely the normativity involved in technological development).

In anti-war novels like *Catch-22* [29] we find individuals who are manipulated by different parties such that they can push them towards particular ends. For example, the anti-hero of Catch-22, Yossarian, is an American bombardier sent to Europe in the Second World War. During his time, he is confronted with the fact that he can never stop flying. His commanding officer always raises the number of flights required to go home each time he gets

close to that magic number. Perhaps if he could be declared unfit for duty, he could stop. However, because of Catch-22, he can never get out. Simply put, the only way to stop flying was to be insane, their only way to show that was to fly. If one were concerned for their safety, they would be sane, and thus had to fly. If one was not, one could fly in the first place. As a result, one had the option to try to excuse themselves from flying, which meant mandatory flights, or one didn't, which resulted in the very same.

The example of Catch-22 clues us into many different facets of control. Control exists between people in power relations. This notion of control is authoritative, it is the way to exert influence on how things go and how people behave. Yossarian cannot control his own actions. Instead, he is forcefully pushed down a particular road. Yet, there is a more general theme in anti-war novels like Catch-22, but also *De Donkere Kamer van Damokles* [30], and *Mother night* [31]. What is repeated over and over throughout these novels is a lack of clarity, which also lets individuals exert control over others. People are heavily misinformed, and the truth is obfuscated to the extent that even those in the know seem to lack the required knowledge. Yet, some plan is executed for reasons unbeknownst to many, and this too is an act of control. This is also where control borders on the sheer dystopic, like George Orwell's *1984* [32] and Bradbury's *Farenheit 451* [33]. In such novels, control over society is exerted in a way that allows it to become an entire lens, a framework of how to think and how to act. Next to the power relation between people, this suggests that there may also be institutional and epistemic kinds of control. The bureaucracy of a state or institute is capable of exerting influence similarly to Yossarian's officer, but so can knowledge exert some truth over us by providing the frame through which we think.

With such novels in mind, we can already see one pillar of control arising: to exert influence over another, or something, is an act. Control and action are heavily linked. There is a relation of control between the commanding officer and Yossarian, in a way that makes Yossarian unable to act freely. The commanding officer is directing the flow of Yossarian's behaviour, and he can only struggle like a dog on a leash. His exertion of control lies in resisting the situation. Fundamentally, control grants us the capacity for action. If we are controlling something, we can steer that thing into a direction we like. Control can be over other people, institutions, but also over concepts like truth, if we look at the anti-war novels I previously mentioned. Control in terms of opposition can thus be thought of as the ability to resist the control of others (and other things)

Yet, that is definitely not the entire picture. In literature about addiction, mental health, and obsession, we see a lack of self-control. In works like: *House of Leaves* [34], *Junky* [35], *The Trial* [36], *One flew over the cuckoo's nest* [37], and *Infinite Jest* [38], we are confronted with various ideas and individuals who lose control, not only because of another human being but also

because of an obsession, an addiction, an institution, medication, and even a film. To take Infinite Jest as an example, one of the themes in this hysterical realist novel is a film which is so addictive that people cannot help but watch it. They forget to do anything else. In this simple way, a loss of control equals a loss of bandwidth in terms of action. Interestingly, the consensus between all the works above is also a measure of character. The ability to resist control or temptation of substances is dependent in part on who we are, what environment we were fostered in, and who we have around us[1].

So, in short, the intuition about control revolves around the capacity to influence things and others in the world according to one's liking. The capacity is an obvious necessity. Is one's liking also quite as obvious? We can be made to influence things even if we do not desire them. Yossarian did fly his plane and had control over it. Yet, he surely wasn't liking it.

How is this reflected in scientific literature? Well, broadly speaking, there are various ways we can talk about control. Nonetheless, the first place to start in philosophy is agency. By necessity, control is closely related to intentional action [26, 40]. If we intend to act in a certain manner, then we have exerted control in order to aim at bringing the current state of affairs about. This does not immediately mean we have the full causal capacity to succeed. Yet, it is intention to act in a certain manner and aim towards a particular end, which ties action and control together. At the most basic level, we would have to discuss free will [41, 42], as it provides the contours for what we could deem as significant control over our actions. If we were without any means to alter our behaviour, beliefs, or actions, or alter our environment in any way, then we can hardly be said to be in control of it.

The paragraph above clues us into a significant subject related to the topics at hand, namely, (moral) responsibility [43–45]. When discussing an agent's control over actions, we should wonder to what extent such an action is also *theirs*, in the sense that they can be held accountable and responsible. The bonds between control and responsibility, are intuitively obvious. Is such an agent the source of their own actions? Or are they willed to that fact, coerced to do so? When dealing with moral responsibility and control, we need to give some account of how one can garner responsibility. If something like free will is out of the question, we are out of luck. Yet, in his influential paper, Harry Frankfurt does pose the problem in quite a salient way [45]. An agent can only be held responsible if said agent could have done otherwise. The question remains, how do we understand *doing otherwise*? One way to go about this is the historicist theory of Fischer and Ravizza [46]. What Fischer and Ravizza provide is a difference between types of control, namely *regulative* and *guidance* control. Thus far, we have only discussed regulative control, which is the intuitive notion: namely, we are capable of doing or not doing an action, and we are capable of doing another action instead. Regulative control

[1] This means that control has some roots also in the Greek concept Akrasia [39]: weakness of will.

has fundamental ties to freedom. For example, we exercise regulative control when we decide to walk off the top of a building, even though it is harmful to us. At least, if that situation also allowed us to not take said action or just take the stairs down instead. Guidance control requires something less. Suppose you are learning to drive a car, and your instructor is next to you. You desire to control the car, but since you are learning, your instructor can intervene. Now you are coming to a turn, so you slow down a bit such that you can make the turn. If, however, you had done something wild, the instructor would have been able to correct you, to make sure the car still behaved properly. Who is in control? Well, the instructor has regulative control, and so long as you behave, you have guidance control. You are freely taking responsibility for an action, and you are guiding the car in a particular direction.

The example of the instructor also clues us in to another dimension of control: *direct* versus *indirect* control. The instructor has delegated control over the car to you, until they decide to intervene. This is where a problem of control obviously arises. You could make a sharp turn before the instructor has time to intervene. In the same vein, Yossarian could fly away. Indirect control opens up the possibility of resistance and thus the complete loss of control. This is where we should talk of the control paradox [47]. This paradox is relatively easy to explain. A password may grant me control over who can get into my e-mail account. At the same time, I can forget it or lose sight of it and thereby lose control over my account myself. Now I don't want to forget stuff like this anymore, so I install a password manager and remember my master key. In doing so, I have centralized all my passwords, I have gained more control, but I also end up risking more. If I forget my master key, I am suddenly unable to access all my account. The paradox of control is that we risk losing more control once we decide to gain more control.

For Fischer and Ravizza, we can say an agent is responsible as long as guidance control comes into existence in the correct way and if an agent has the right attitude towards the actions. This attitude comes along once the outcome of an action is decided by a mechanism which is *moderately reason-responsive*. We can say that someone having guidance control is responsible if they also have a responsibility attitude towards it. What this all entails is that there is a mechanism (some way the action came about) that is moderately dependent on reasons. Meaning that given a different set of reasons, one thing would have led to a different action. We can further delineate between reason-responsiveness by suggesting that it ought to be: reason-receptive and reason-reactive. The mechanism is reason-receptive insofar as it depends on the cognitive skills of the agent to understand the implications of actions given. The mechanism is reason-reactive if it gives rise to different actions if the mechanism is fed different reasons.

So, we have seen some connections between control and responsibility, but how does this give rise to questions of technology? Well, the question

of control in systems actually came from somewhere else. The fields of cybernetics and control theory have rooted debates in terms of control as well. Wiener, one of the grandfathers of AI, for example, talks about Cybernetics as the control and communication of animals and machines [48]. It was a system agnostic way of discussing information processing. Control has a very different meaning here, rather than being about the control of actions, it is a governor [49] or controller which stabilizes a process to minimize fluctuations. The observer plays a much more passive role at first glance, as we try to automatize a process for which, another unit - a governor - can exercise control over a system. One easy example that you may encounter in daily life is a proportional-integral-derivative controller (PID), which you can find in a good espresso machine. A PID is used to adjust the temperature to maintain it at the right level for coffee. It means that it can stabilize the fluctuations of temperature by giving outputs based on input. It contains a feedback loop, which is exactly one of the fundamentals of cybernetics and control theory. Controllability is not necessarily about the control of our actions, but rather the influence we can have on the irregularities of a system through another module (a controller).

The similarities between Fischer and Ravizza's moderately reason-responsiveness and controllability are interesting. In both cases there is input, on which a controller (either we or another unit) proposes a signal. Reasons and output are different words for a rather similar situation. The difference between what Fischer and Ravizza propose as compared to control theory is an ability to act on different signals. A PID-controller in an espresso machine will likely not have means of interacting with the signals of video footage of a self-driving car, while we might be able to interact with both. Second-order cybernetics, which includes the observer into the loop [50] of feedback mechanisms, rather than disguising them [51], seems more appropriate to discuss the current climate of technological innovations. It would posit certain AI systems as controller units that provide a stable output based on input. It becomes a question whether the observer is within the system over which a controller exerts output, or whether it is actually an observer.

The theory of Fischer and Ravizza has also given rise to a particular version of meaningful human control in philosophy of technology [16, 17]. Originally meant for understanding responsibility with regard to autonomous weapon systems, the concept has been taken up in a larger sphere to discuss how we ought to think about control and technology. The concept applies much of Fischer and Ravizza's theory practically. The baseline is this: human beings ought to be controlling machinery at the end of the day, but must do so in a meaningful way. Thus, the mechanism which is moderately reason-responsive is poured into a model which allows us to consider who was in (guidance) control in a given situation such that they can also be held responsible. The reason-receptivity and reason-reactivity, are boiled down to *trac-*

king and *tracing*. The concept of tracking is the necessity of the system to keep track of the relevant moral reasons in the given situation. Ergo, if a system is meant to keep cold-blooded creatures alive, then it is important to track the temperature. Tracking is the topic of chapter 3. As we will see, tracking itself is highly intractable, and will require additional requirements to actually make sense of what it means. It appears that tracking requires a particular goal to know what to track accordingly, and this is exactly one of the issues that has been problematic within computer science. Knowing what is relevant to a given situation and/or goal.

Moving into the territory of *tracing*, we find that it is a co-authorship between system and humans. There should be a traceable lineage to a (set of) human actor(s) and a decision process such that said actor(s) can be held accountable. It requires the reason-receptivity of Fischer and Ravizza by virtue of the fact that responsibility can only be attributed when an agent has the capacity to properly understand the implications of said actions. The problems surrounding making tracing actionable is much of what we will discuss in chapter 4 and 5. Given that we design systems in a particular way, it seems to require that users either have a serious grasp of the implications of an action or designers know an immense amount about the specificity of the user. Since users differ wildly in cognitive capacity, environment, and understanding of certain notions, we cannot meaningfully attribute responsibility to either party without attributing some notion of intelligence to said human agents, which is likely extraordinary. We will use the probability of Personal Artificial Intelligent Assistants to show how both a lack of definition of user and capacity creates a conundrum for taking on a responsibility attitude towards such technology.

Control and responsibility are interlinked by virtue of knowing that actions are one's own, thus being able to effectively take the blame or praise when one guides a system. This is, in the current climate, often problematic when we have technology in mind [6, 52]. Are we truly responsible for the actions of a machine? Or do they possess enough "agency"to be considered blame- and praiseworthy themselves? The concepts of *responsibility gaps* and *achievement gaps* [53] exist in the interplay between humans and technological systems, in which it isn't immediately obvious who ought to be held responsible. In some sense, it is the problem of many hands [54], these systems are designed by individuals, and their design in some way carries part of the responsibility for action. If we think about it in terms of tracking and tracing, they designated what ought to be tracked, but in general, design choices can lead to situations where nobody really seemingly desires to carry blame. For example, mistakes made by a self-driving car on which an individual relies have a sort of shared blame between driver and car, but who is ultimately held responsible? These sorts of questions matter, both for accountability and also to ensure some kind of humanity in society. Life could turn out brutish, na-

sty, and short, if people can get hurt or worse without ever having someone to account for it. It incentivizes carelessness and neglect.

That brings us to the larger picture. We may be talking about the responsibility of a particular machine, but there is also an element of both control and responsibility in the larger discourse and shape technologies take. In another aspect of control, we need to discuss the particular understanding of what technology can do to our sensemaking. There is perhaps no more pressing dilemma in the philosophy of control than the Collingridge Dilemma [11]. As mentioned in the previous section, there appears to be a lack of oversight about the consequences of technology. This, too, is a question of control. If we do not know what the consequences may be, how can we meaningfully exert influence over its design and move society towards better technology? The ties to knowledge come into play once we understand a bit more about *mediation theory*.

Mediation theory is a way to understand our relationship with technology [13]. What it boils down to is a particular understanding that technology shapes our relationship to the world and each other. Mediation plays an active part in the constitutive factors of being and of knowledge. Once we invented the way to look at a baby in a belly, said baby also had the potential to become a patient. We could talk about it in new ways, relate to it differently. If it were sick, for example. Technology and knowledge are intertwined in this sense because it can provide a lens through which we can view things. If we hearken back to the start of this section, be reminded of 1984. Technology is in and of itself a lens, a framework of how to think and act. This is also heavily implied by *technological affordances* [55], which are the designs that elicit certain responses. Consider, for example, a push bar on a door vis-à-vis a door handle. One elicits pushing, the other a certain motion of the hand. As a colleague put it, a fork is designed to elicit eating; an atomic bomb is (hopefully) not. The relationship between the Collingridge Dilemma and mediation theory is what necessarily sets up some antagonistic relation to control. This is discussed in chapter 2. Specifically, chapter 2 deals with the notion of path-dependence [56, 57], namely that technology often opens pathways to new technology and closes others. This obviously may make it difficult to consider alternative pieces of technology if those pathways are simply not open to us in society.

The basics of control, intuitively, boil down to acting freely and influencing the world. Yet when we come to technology, the ability to actively choose different alternative actions is limited. Perhaps because some control is taken out of our hands by an actual controller, or perhaps because it elicits a particular action over another. What does that really mean for our responsibility when talking about human-technology systems? The design and implementation of technological devices influences society at large. There too seems to lie a kind of control and responsibility that is not often talked about. Are

these not considerations that are fundamental not only to the debate about technology but also to society and its movement towards the future?

1.3. ON WICKEDNESS

Such questions about society and where it goes fit rather neatly with the notion of wickedness. The idea of wickedness stems from Rittel and Weber's seminal paper [58]. The question of wicked problems is juxtaposed with tame problems [59]. An easy way of understanding wickedness is through tame problems. A good example of those types of problems are the ones computer science students used to get during their studies: Sudoku solvers and map colouring algorithms. Such problems are tame because there is one obvious single correct answer: namely a correctly solved sudoku and a fully coloured map (in which none of the adjacent squares share a colour). Such tame problems have part of their origin in Herbert Simons [60].

What exactly is a wicked problem? Well, according to Rittel and Weber, a wicked problem:

- Has no clear problem definition.
- Can never be solved completely.
- Does not have solutions which are right or wrong, only better or worse.
- Does not have an ultimate test to see whether it is solved.
- Can only be solved in ways that have ramifications for the entire system it is in.
- Has no exhaustively describable set of potential solutions.
- Is essentially unique.
- Is always connected to other wicked problems.
- Is not solved by scientific inquiry.
- Has impact on the lives of people.

As a consequence, wicked problems are related to control. If wicked problems are dependent on our definition of them, then those who are designing technology are also the ones who are applying a certain solution. Taken together with the fact that there is always more information to be gathered, and we are without sufficient means to solve it completely, means certain limitations to control apply. For example, a solution and problem formulation may not be ours to begin with. It could be that our concerns are not taken into account in the first place. Furthermore, if we are without means to test whether it is solved, we may accept a solution that is insufficient for us. In any case, wickedness stands in relation to control because it asks us what we can and cannot control.

1.4. WHAT THIS DISSERTATION IS ABOUT

Let's get back to the example of the introduction: we discussed a man who could eavesdrop on his ex through a wall-mounted tablet. Why did this happen? The man could act freely in an undesirable way. If we think about abuse of smart appliances, namely turning the lights on and off, we clearly see a mistake in our technological development but also deployment. One is using a device and exerting control, and the other needs to resist it. Yet, surely societal factors and commercial factors were wrapped up in this problem as well. If governance had prohibited the use of off-the-shelf parts for such a situation, then it would have been unlikely that such a wall-mounted tablet had a microphone in the first place. Control of both deployment, creation, and use, seems shared between a multitude of agents. The problem of misuse boils down to a combination of control and wrongful deployment or development. But what is necessary to prevent such a thing?

The man was clearly in control of the tablet and can be held accountable for his actions, but what about the fact that there was a microphone present in the first place? Perhaps the man could have intentionally bought it at that time, but it is far more likely he only found out after the fact. Does it matter that we invented microphones in the first place? Does it matter that designers may have overlooked this fact or did not consider it during deployment? All of these questions invite the thought that such an invention may have been a solution to a wicked problem. Does it matter that technology alters our ways of living in our own homes? Does it matter that our ability to exert control over another is being altered through such means?

Technological development is wicked and normative and carries a cost. To solve issues, it is essential that we look at development through a philosophical and political lens.

That statement sounds uncontroversial, yet it is those costs we often overlook. It is rather easy to be mindful of the benefits of technology. Consider, for example, that modern technology helps us to live longer and allows us to travel further. The improvements in modern-day life can make it seem as if technology is almost always a net positive. It may be difficult to see the negatives without also undoing the beauty that has arisen alongside the changes. While some may argue in favour of blunt technological progress, I would rather posit that technological development is always a disruption of standards of living. One way of life must die for another to exist. Horse and carriage being trumped by cars, it becomes hard to live in a world where horses are the dominant form of travel. We may gain speed by traveling in cars, allowing ourselves to cross great distances to see family and friends. Yet, we also lose something else. Think of the centralization caused by cars, as the distance crossed in hours is far larger than ever before. Because of it, we require fewer villages and inns along the road and less of an infrastructure to support

travelers. While we could start using cars at a slower pace, their capacity for speed also implies an understanding of it. This is not in and of itself a negative. Standards do need to change, but we should sincerely consider what costs are desirable and acceptable in our society.

The wickedness comes in due to the fact that we are developing and deploying technology in a situation where we cannot know all the relevant patterns, and we do not know what we need to know to solve it correctly, and likely it is heavily dependent on how we frame the question in the first place. If we want to solve the issue of transportation, and we consider speed to be a major factor, then travel by planes may be worthwhile to pursue. If we consider the climate to be a relevant factor instead, we might avoid planes altogether.

Yet, we live in a world in which technological development is often heralded as progress. There is enough of a debate as to what the downsides are of particular kinds of technology. As well as many of the worries we have about particular devices. Much of such conversations are after the fact, once it has been let loose in the world. This is in no small part because we often do not really understand what technology will do once it is out in the wild. While there are ways like technological assessment [61], and notions of sociotechnical experiments [62] or technomoral scenarios [63], we need a clearer and more ubiquitous standard by which we can measure the worth of a device, or otherwise, we truly need ways to counteract and destroy technologies which are altering our world for the worse. It is not only that these things are normative, but we also require understanding in what way they are normative. So this leads me to the research question: **What are the costs of technological development, and what control can we exert over these costs?**

1.5. BACKGROUND

Broadly speaking, this entire dissertation should be set against the debate of meaningful human control. Even though value alignment and responsibility gaps are allied concepts to that debate and thus are in turn mentioned and discussed within the work, and the entire dissertation is about costs of technological development, this is all done within the scope of meaningful human control and finding means to properly foster the environment such that this method is effective and worthwhile to pursue. The stance within the dissertation has a political focus surrounding meaningful human control, which is novel, and the problems discussed help move the debate surrounding its use forward.

To briefly mention the allied concepts, the notion of responsibility gaps deals with the lack of obviously attributable responsibility to individuals in situations where we deal with autonomous agents. Value alignment is an attempt to avoid mishaps in the first place by trying to make sure such systems act appropriately. You can find an overview of these and many other different concepts discussed within this dissertation in appendix A.

In terms of meaningful human control, there is some research being done as to how to apply it both generally and in more specific domains [16, 18, 19, 64–67]. The reason for mentioning both value-alignment and responsibility gaps is because meaningful human control is a solution to both at the same time. The main way I look at meaningful human control is through Santoni de Sio and Van den Hoven [16], and in their case, meaningful human control is meant to track the relevant moral reason within a context and to trace it to some set of individuals who can be held morally responsible. The first is an attempt to solve some of the misalignment with our goals; the latter is an attempt to attribute responsibility where it is due.

So what gap does this dissertation fill? I use meaningful human control as a means to examine the costs that come with technological development even if we try to do it well, ergo if we desire to avoid misalignment and try to attribute responsibility where it ought to be placed. Especially when glanced at through a political lens, we still see that issues and costs may arise, and to counteract those seems to be well beyond the capacity of what meaningful human control can provide us. Thus, it is a way of gesturing towards the limits of such a theory in general and also to provide requirements that may lie relatively outside the boundaries of meaningful human control but are nonetheless necessary for it to be achieved.

There are generally speaking philosophical perspectives [16, 18, 52, 67, 68], legal perspectives [19, 69–71], and more technical perspectives [65, 72–74] but there is rather little about the political problems that may still arise from the technologically disenfranchised, minorities, and the problems of power that may all co-exist with these kinds of technologies even if we do try to make use of meaningful human control. It is this gap that the dissertation tries to fill. If you are wondering what this has to do with the costs of technological development, then you can understand meaningful human control simply as a means of trying to implement a system properly.

While there has been a recent upsurge in political theory surrounding AI systems [75, 76], the debate is only starting to get excavated. Of course, there are a lot of older works on society and technology [11, 12, 24, 77–79], which do inform the work you currently have in your hands.

1.6. OUTLINE

Now that we have gone over the basics, we can delve into the particularities of the chapters.

In Chapter 2, I examine the issue of wickedness through self-reinforcement. If we desire control over technology, and especially to counteract some of the costs of control, then we need to understand that we often do not know the consequences of technology before their implementation. Yet, they do influence both our society and our moral frameworks. It also provides the start for new kinds of technologies. The problem therein lies, that good

technology may form the basis for terrible technology down the line. This is an obvious issue for control, and it has much to do with the fact that we have no premonition of what technology may entail. In other words, we may be able to steer technology innovation, it may just not be clear which haven it leads us to.

In Chapter 3, I will go over the aspect of relevancy. If we want to control technology both from a user perspective and a design perspective, we want to properly provide the right teleological function, the right data, and the right possible patterns of inference. In other words, we need to get the relevancy right. However, once again we have to contend with our lack of premonitions. We do not know what may be relevant in a given situation, and we are hard-pressed to make sure all the relevant details are relevantly portrayed during design. This causes serious issues for control, as users may not be fully portrayed. To solve such issues, we need to accept that an iterative cycle is the best solution we have at hand, even though that is debatable in classical thought surrounding wickedness.

In Chapter 4, I look at Personal Artificial Intelligent Assistants and concern myself with the notion of a leg up for users as a way of giving a more detailed account of what may go wrong once we think technology tries to improve our lives. The problem we pose is that such enhancement is unequally distributed and perhaps veering in the wrong direction. Even if we desire to create design requirements through something like meaningful human control, we require that such artefacts are effectively theorized so we know what they are capable of and why.

In Chapter 5, I demonstrate that tracing within meaningful human control has the goal to attribute moral responsibility, but the theories on which it is based do not give a proper account of how that should happen. The flavours of moral responsibility attribution do not hold themselves up to scrutiny, leading to the fact that we have an issue with correctly establishing responsibility. I suggest that this can be solved by introducing a normative dimension with responsibility attribution, building on Young's social connection model of responsibility as justice [80].

What these chapters should show together is why it is so difficult to understand the costs of technology. We are met with lofty promises (see Chapter 4) but those realities cannot be met once we try to integrate them into society. This is simply the case because of self-reinforcement (see Chapter 2) and a lack of adequate knowledge (see Chapter 3). One way forward would be a radical shift in how we attribute responsibility when it comes to design (see Chapter 5). But overall, we must also consider that the way we think about technology itself has to be subject to the topic of technological deployment, which is what I turn to in the discussion (see Chapter 6).

BIBLIOGRAFIE

[1] A. Riley, "How your smart home devices can be turned against you", *BBC Future*, 2020.

[2] R. Leitão, "Anticipating smart home security and privacy threats with survivors of intimate partner abuse", in *Proceedings of the 2019 on designing interactive systems conference*, 2019, p. 527–539.

[3] D. McKay en C. Miller, "Standing in the way of control: A call to action to prevent abuse through better design of smart technologies", in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, p. 1–14.

[4] J. A. Naslund, A. Bondre, J. Torous en K. A. Aschbrenner, "Social media and mental health: Benefits, risks, and opportunities for research and practice", *Journal of technology in behavioral science*, jrg. 5, p. 245–257, 2020.

[5] R. Benjamin, *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons, 2019.

[6] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and information technology*, jrg. 6, nr. 3, p. 175–183, 2004.

[7] D. W. Tigard, "There is no techno-responsibility gap", *Philosophy & Technology*, jrg. 34, nr. 3, p. 589–607, 2021.

[8] T. J. Gee, "Why female sex robots are more dangerous than you think", *The Telegraph*, jrg. 5, 2017.

[9] T. Kubes, "New materialist perspectives on sex robots. a feminist dystopia/utopia?", *Social Sciences*, jrg. 8, nr. 8, p. 224, 2019.

[10] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[11] D. Collingridge, "The social control of technology", 1982.

[12] M. Kranzberg, "Technology and history: ''kranzberg's laws''", *Technology and Culture*, jrg. 27, nr. 3, p. 544–560, 1986, ISSN: 0040165X, 10973729.

[13] P.-P. Verbeek, "Toward a theory of technological mediation", *Technoscience and postphenomenology: The Manhattan papers*, jrg. 189, 2015.

[14] P. Vermaas, P. Kroes, I. Van de Poel, M. Franssen en W. Houkes, "Technical artefacts", in *A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems*, Springer, 2011, p. 5–20.

[15] I. Gabriel, "Artificial intelligence, values, and alignment", *Minds and machines*, jrg. 30, nr. 3, p. 411–437, 2020.

[16] F. Santoni de Sio en J. Van den Hoven, "Meaningful human control over autonomous systems: A philosophical account", *Frontiers in Robotics and AI*, jrg. 5, p. 323 836, 2018.

[17] L. C. Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker e.a., "Meaningful human control over ai systems: Beyond talking the talk", *arXiv preprint arXiv:2112.01298*, 2021.

[18] H. Veluwenkamp, "Reasons for meaningful human control", *Ethics and Information Technology*, jrg. 24, nr. 4, p. 51, 2022.

[19] J. Davidovic, "On the purpose of meaningful human control of ai", *Frontiers in big data*, jrg. 5, p. 1 017 677, 2023.

[20] N. Cornelissen, R. van Eerdt, H. Schraffenberger en W. F. Haselager, "Reflection machines: Increasing meaningful human control over decision support systems", *Ethics and Information Technology*, jrg. 24, nr. 2, p. 19, 2022.

[21] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Rios, J. Bobes-Bascarán en Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art", *Artificial Intelligence Review*, jrg. 56, nr. 4, p. 3005–3054, 2023.

[22] B. Friedman, P. H. Kahn, A. Borning en A. Huldtgren, "Value sensitive design and information systems", in, Springer, 2013, p. 55–95.

[23] D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell en A. Dragan, "Inverse reward design", *arXiv preprint arXiv:1711.02827*, 2017.

[24] A. Feenberg, *Questioning technology*. Routledge, 2012.

[25] K. Leins, H. Arendt en J. Weizenbaum, "Ai for better or for worse, or ai at all", *Future Leaders*, 2019.

[26] J. Shepherd, "The contours of control", *Philosophical Studies*, jrg. 170, p. 395–411, 2014.

[27] J. M. Fischer en M. Ravizza, *Responsibility and control: A theory of moral responsibility*. Cambridge university press, 1998.

[28] M. Foucault, "Discipline and punish", in *Social theory re-wired*, Routledge, 2023, p. 291–299.

[29] J. Heller, *Catch 22*. Grasset, 2004.

[30] W. F. Hermans, *De donkere kamer van Damokles*. Bezige Bij bv, Uitgeverij De, 2012.

[31] K. Vonnegut, *Mother night: A novel*. Dial Press Trade Paperback, 1999.

[32] G. Orwell en E. Fromm, *1984: A Novel*, reeks A Signet classic. Signet Classics, 1977, ISBN: 9780812416299. adres: https://books.google.nl/books?id=4HQcMQAACAAJ.

[33] R. Bradbury, *Fahrenheit 451: A Novel*. Simon & Schuster, 2011, ISBN: 9781439142677. adres: https://books.google.nl/books?id=OYtkbGl2j0sC.

[34] M. Danielewski, *House of Leaves: The Remastered Full-Color Edition*. Knopf Doubleday Publishing Group, 2000, ISBN: 9780375703768. adres: https://books.google.nl/books?id=o6NzGNODRjkC.

[35] W. Burroughs en O. Harris, *Junky*, reeks Penguin Modern Classics. Penguin Books Limited, 2012, ISBN: 9780141904016. adres: https://books.google.nl/books?id=2CvEs2xybZ4C.

[36] F. Kafka, *Der prozess: roman*. Suhrkamp Verlag, 2012.

[37] K. Kesey, *One Flew Over the Cuckoo's Nest: (Penguin Orange Collection)*, reeks Penguin Orange Collection. Penguin Publishing Group, 2016, ISBN: 9780143129516. adres: https://books.google.nl/books?id=KhRvDQAAQBAJ.

[38] D. Wallace, *Infinite Jest: A Novel*, reeks An Abacus Book: Fiction. Abacus, 1997, ISBN: 9780349108773. adres: https://books.google.nl/books?id=3gxhQgAACAAJ.

[39] P. A. David, "Path dependence, its critics, and the quest for 'historical economics'", in *Evolution and path dependence in economic ideas*, Edward Elgar Publishing, 2001, p. 15–40.

[40] G. Sher, "Out of control", *Ethics*, jrg. 116, nr. 2, p. 285–301, 2006.

[41] N. v. Miltenburg, "Freedom in action", proefschrift, Utrecht University, 2015.

[42] A. Feltz en F. Cova, "Moral responsibility and free will: A meta-analysis", *Consciousness and cognition*, jrg. 30, p. 234–246, 2014.

[43] J. M. Fischer, "Recent work on moral responsibility", *Ethics*, jrg. 110, nr. 1, p. 93–139, 1999.

[44] M. Talbert, *Moral responsibility: an introduction*. John Wiley & Sons, 2016.

[45] H. Frankfurt, "Alternate possibilities and moral responsibility", in *Moral responsibility and alternative possibilities*, Routledge, 2018, p. 17–25.

[46] J. M. Fischer en M. Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, reeks Cambridge Studies in Philosophy and Law. Cambridge University Press, 1998.

[47] E. Di Nucci, *The control paradox: From AI to populism*. Rowman & Littlefield, 2020.

[48] N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

[49] J. C. Maxwell, "I. on governors", *Proceedings of the Royal Society of London*, nr. 16, p. 270–283, 1868.

[50] D. A. Novikov, *Cybernetics: from past to future*. Springer, 2015, deel 47.

[51] B. Scott, "Second-order cybernetics: An historical introduction", *Kybernetes*, jrg. 33, nr. 9/10, p. 1365–1378, 2004.

[52] S. Nyholm, "Responsibility gaps, value alignment, and meaningful human control over artificial intelligence", in *Risk and responsibility in context*, Routledge, 2023, p. 191–213.

[53] J. Danaher en S. Nyholm, "Automation, work and the achievement gap", *AI and Ethics*, jrg. 1, nr. 3, p. 227–237, 2021.

[54] I. Van de Poel, "The problem of many hands", in *Moral responsibility and the problem of many hands*, Routledge, 2015, p. 50–92.

[55] W. W. Gaver, "Technology affordances", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1991, p. 79–84.

[56] S. E. Page e.a., "Path dependence", *Quarterly Journal of Political Science*, jrg. 1, nr. 1, p. 87–115, 2006.

[57] P. A. David, "Clio and the economics of qwerty", *The American economic review*, jrg. 75, nr. 2, p. 332–337, 1985.

[58] H. W. Rittel en M. M. Webber, "Dilemmas in a general theory of planning", *Policy sciences*, jrg. 4, nr. 2, p. 155–169, 1973.

[59] R. Coyne, "Wicked problems revisited", *Design studies*, jrg. 26, nr. 1, p. 5–17, 2005.

[60] H. A. Simon, *The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird*. MIT press, 2019.

[61] D. Banta, "What is technology assessment?", *International journal of technology assessment in health care*, jrg. 25, nr. S1, p. 7–9, 2009.

[62] I. Van de Poel, "Society as a laboratory to experiment with new technologies", in *Embedding New Technologies Into Society*, Jenny Stanford Publishing, 2017, p. 61–87.

[63] T. Swierstra, D. Stemerding en M. Boenink, "Exploring techno-moral change: The case of the obesitypill", *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments.*, p. 119–138, 2009.

[64] G. Mecacci en F. Santoni de Sio, "Meaningful human control as reason-responsiveness: The case of dual-mode vehicles", *Ethics and Information Technology*, jrg. 22, nr. 2, p. 103–115, 2020.

[65] L. Cavalcante Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker e.a., "Meaningful human control: Actionable properties for ai system development", *AI and Ethics*, jrg. 3, nr. 1, p. 241–255, 2023.

[66] F. S. de Sio, G. Mecacci, S. Calvert, D. Heikoop, M. Hagenzieker en B. van Arem, "Realising meaningful human control over automated driving systems: A multidisciplinary approach", *Minds and machines*, jrg. 33, nr. 4, p. 587–611, 2023.

[67] S. Robbins, "The many meanings of meaningful human control", *AI and Ethics*, jrg. 4, nr. 4, p. 1377–1388, 2024.

[68] S. Nyholm, "Automated mobility and meaningful human control: When and why is control important and what are its different dimensions?", in *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, Edward Elgar Publishing, 2024, p. 13–27.

[69] R. Crootof, "A meaningful floor for meaningful human control", *Temp. Int'l & Comp. LJ*, jrg. 30, p. 53, 2016.

[70] B. Boutin en T. Woodcock, "Aspects of realizing (meaningful) human control: A legal perspective", in *Research Handbook on Warfare and Artificial Intelligence*, Edward Elgar Publishing, 2024, p. 179–196.

[71] G. Contissa, "Legal and governance perspectives on meaningful human control: The case of automated mobility", in *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, Edward Elgar Publishing, 2024, p. 82–103.

[72] S. C. Calvert, D. D. Heikoop, G. Mecacci en B. Van Arem, "A human centric framework for the analysis of automated driving systems based on meaningful human control", *TheoreTical issues in ergonomics science*, jrg. 21, nr. 4, p. 478–506, 2020.

[73] J. Kwik, "A practicable operationalisation of meaningful human control", *Laws*, jrg. 11, nr. 3, p. 43, 2022.

[74] J. van Diggelen, K. van den Bosch, M. Neerincx en M. Steen, "Designing for meaningful human control in military human-machine teams", in *Research handbook on meaningful human control of artificial intelligence systems*, Edward Elgar Publishing, 2024, p. 232–252.

[75] M. Coeckelbergh, *The political philosophy of AI: an introduction*. John Wiley & Sons, 2022.

[76] S. Zuboff, "The age of surveillance capitalism", in *Social Theory Re-Wired*, Routledge, 2023, p. 203–213.

[77] B. Latour, *Aramis, or the Love of Technology*. Harvard University Press, 1996.

[78] H. Tsoukas, "The tyranny of light: The temptations and the paradoxes of the information society", *Futures*, jrg. 29, nr. 9, p. 827–843, 1997.

[79] J. Ellul, *The technological society*. Vintage, 2021.

[80] I. M. Young, *Responsibility for justice*. Oxford University Press, 2011.

2

CAN WE CHANGE OUR AI SYSTEMS?

If the doors of perception were cleansed everything would appear to man as it is, infinite. For man has closed himself up, till he sees all things through narrow chinks of his cavern.

William Blake, *The Marriage of Heaven and Hell*

The Collingridge Dilemma states that once technology is embedded, it becomes very hard to change. If we want to have a better implementation, then we need to be able to alleviate some of the burdens of embedding. In this chapter, I describe how both path-dependency and technomoral change have to deal with self-reinforcing tendencies of technology. As our perception is altered by the introduction of technology, this alteration may lead us down particular paths of exploration with less and less deviation. If we accept any kind of exploration as worthwhile, we run the risk of inviting moral relativism. To combat this, I argue in favour of interoperable technology with limited scalability. This makes it so that technology can be swapped if it proves problematic, and that the impact of technology is dampened. Aside from this, we also need to attend to the fact that such an implementation is going to have an impact on the way we perceive certain things, even if it is an unintended side effect. As such, I argue that there is a mediative responsibility which resides with designers, precisely because they chose to design something in the first place.

2.1. INTRODUCTION

Technology has, without a doubt, a major impact on our species and our society [81]. Technology is sometimes heralded for its efficiency, for example, bureaucratic systems can become more responsive to citizens' needs, cheaper, and efficient [82]. Indeed, in the current day, we create artificial intelligence (AI) systems which have the capacity to structure huge amounts of data or make an incredible number of inferences in little time. But long before AI, we had debates about the worth of trains, steam engines, and coal use [83, 84]. Electricity has profoundly changed our relationship to one another and our lives in almost every aspect. We can live longer and we can travel farther. Yet, technology rarely seems to come along without downsides. Mechanised production of goods gave rise to worse working conditions. The invention of the combustion engine is quickly destroying the planet with CO2 emissions. The means for invading our privacy and surveillance is greatly enhanced by cameras but also by AI [76]. The capacity of ethnic profiling is greatly increased too.

One of the major questions that comes along with technology is the question of embedding in society. Classically, this follows from The Collingridge Dilemma [11]. The dilemma states that we may be able to influence technology early on, even though we do not know what impact it will have. Yet, once technology has become embedded, it becomes difficult to remove and influence. This kind of entrenchment has explicit ties to path-dependency [85]. Not only is technology entrenched, it also creates a path that is difficult to deviate from, such that new technology can deviate less from the beaten path. If we take this into account, together with ideas of technomoral change [63, 81, 86, 87], then we see a problem arise. Our moral frameworks adapt through our interaction with technology and because our frameworks and perspective change, so too changes our ideas of what we can in turn create or invent. Not only is the entrenchment of technology a way in which new and very different technology is harder to conceive of, it also makes it inherently more difficult to consider alternative moral frameworks. This interaction between experience and embedding is already debated in the corner of environmental sciences [88]. In that debate, it is called the extinction of experience, as the progressive loss of human-nature interaction (partially because of technology) changes our perspective on nature and may hamper the support for probiodiversity policies [89], which in turn reinforces the loss of human-nature interaction.

In this chapter, I look at entrenchment as mentioned by The Collingridge Dilemma, through the lens of path-dependency and technomoral change. Path-dependency drives us down a particular path, which is different from merely becoming entrenched. When we lose the ability to conceive of alternative strategies and worlds to live in, we also make technology more of a noose than it needs to be. To take this into consideration, I will first go over a few

strategies that try to deal with the problem of embedding and discuss how AI technology has more trouble with embedding compared to other technologies. I show how they are similar and how longer-term issues arise from that. The major issue is that like the extinction of experience, it is very possible that we lose out on certain moral frameworks due to the self-reinforcing nature of certain artefacts and phenomena, which leads to a similar loss as described by the extinction of experience. Afterwards, I discuss three different options that may address self-reinforcement to a certain extent: interoperability and limited scalability, and better responsibility attribution.

2.2. PROBLEMS OF EMBEDDING

The issue of the Collingridge Dilemma is a supposed double bind. On the one hand, we cannot know the consequences of technology and on the other hand, we cannot influence the technology once we know the consequences, once it is embedded. Meaning that we may be able to control and influence technology at a point where it is not yet meaningful to do so, but when we do want it, it becomes slow and difficult [90]. The embedding of technology leads to varieties of the same issue: entrenchment, path-dependency [39], vendor lock-in [91]. All of which describe technology that is difficult to move away from once embedded.

This in turn leads to the need for corrigibility [90]. Such a need is a search for how we can meaningfully control the flow of technology use and the extensions of innovations. At this need for corrigibility, we can see that two different discussions meet. On the one hand, there is a debate about path-dependency, which is not merely limited to technology but also considers institutions and organizations [92, 93]. And on the other hand, there is a debate about how to steer or deal with technomoral change [81, 94]. They are similar with respect to what they drive at, namely the need for some measure of control in terms of what technology does to us. Within path-dependency, it is a question of what technology follows from previous technologies, or what social effects are derived from it. In steering technomoral change, we are mostly concerned with how our moral frameworks are altered by the onset of technology, which considering the embedding of such technology may give rise and space for new technologies to develop.

2.2.1. PATH-DEPENDENCY

Let us first glance at path-dependency. In its weaker version, path-dependency is the idea that events in the past have an effect on events occurring at a later time [95]. The idea of path-dependency has been portrayed quite neatly by Schreyögg and Sydow [93]. They describe three phases: preformation, formation, and lock-in. The first is an open exploration; the second is a given set of choices made, making it harder to reverse the course of action, but there remains a certain bandwidth; and the third is lock-in,

which means the critical point has been passed, and the path has becomes quasi-deterministic. In institutional and organizational situations, the lock-in phase, is complex and ambiguous in nature, the practices are dependent upon the path before them. The question is how to open up the quasi-deterministic route to some alternative?

Aside from sidestepping the problem altogether (by suggesting it doesn't exist, for example), we can roughly see three versions of solutions: *upfront*, *during*, and *afterwards*. In the first, the policy-maker takes stock of what minor changes may have long-term consequences. Optimistically, Levin et al. [96] suggests that we should fight fire with fire. In other words, we should create policy that kick-starts such self-reinforcing effects which also consider these long-term problems[1]. This suggests that a policymaker should also address those things that lie outside the mainstream political analysis and ask how to create a policy that creates a self-reinforcing effect with such longer-term consequences in mind, but also to ask whether we can create policy that allows for more of a layering effect. In other words, can designers create policy that entrenches itself and allows for more policy that paves that given route?

A perhaps more manageable route is accepting that such solutions aren't complete, and thus, in the case of policy, try to make a short iterative cycle in which one can adapt quickly. This means we try to wrest free from path-dependency during the process. This is much of what Lindblom already portrayed [97]. Small steps make the process of moving towards the right solution more achievable. Of course, it begs the question whether the larger picture is accurately taken into consideration. The idea is to make sure one can regain different paths of exploration during the process itself, thereby allowing for shifts. Of course, we cannot get an overview of all the possible consequences, but instead we focus on the evolution of the problem and address the underlying self-reinforcing patterns in order to regain scope and the ability to manoeuvre towards the solution we desire [98]. In the last option, trying to solve the issue afterwards, we see another kind of iteration cycle. Yet, the emphasis is placed on how one ought to change, rather than how Lindblom suggests an initial limited scope. Fortwengel and Keller suggest interacting with the underlying self-reinforcing tendencies of the structure and addressing them properly can give agency through understanding and interjecting the possibility of evolution. The problem here is the issue of compounding effects. One may regain space to manoeuvre, but is that sufficient space? The consequences of these solutions, may be a lack of feedback, simply assuming enough information at the start (which is never the case for such wicked problems) or a lack of overview, considering the situation at hand and trying something, which acknowledges our limited knowledge. Yet, this begs the question: on

[1] Like the self-reinforcing effect, Lazarus also classifies increasing returns, positive feedback, and lock-in as mechanisms of reproduction which creates this kind of path dependency. The reader can look at the paper *Playing it Forward* [96] for further explanation of the mechanisms that underlie such processes.

which part of our limited knowledge do we act?

2.2.2. TECHNOMORAL CHANGE

Turning to technomoral change, which can be briefly understood as the idea that our perspective on things change through the use of technology. In this debate, we view the Collingridge Dilemma and the problem of control more in terms of how to address the softer side of things. Can we steer technology into a more amendable direction? Here too we see similar solutions arise: *anticipation* [63, 99], *experiments* [100–102], and *mediation analysis* [87]. Starting with *anticipation*, one goes into what the consequences might be and speculate. In one way this is done by anticipation through regulation [99], the idea is to have pre-emptive regulatory power over the innovation process. Another way of envisioning it is through technomoral scenarios [63]. This is to anticipate changes in the ill-defined or softer side of things. Norms will be contested through technology, and such scenarios can help anticipate the social and cultural implications of technologies. In anticipation, we run into the issue that it is a debate about what *the good life* is. What ought to be kickstarted and what not? If we take for example a cue from history, the industrial revolution rapidly changed Great Britain, but in doing so, the quality of products deteriorated rapidly too [103]. Economic output is easily to define, quality is not. Furthermore, there is speculation involved, which makes it difficult to know whether such anticipation is actually right in its approach.

The idea of *sociotechnical experiments*, as described by Van de Poel [100], acknowledges that we cannot oversee everything, but that should not stop us from experimenting with technology in a reasonable way. What this means is that we can experiment on technology in a similar vein as we do with many other experiments, the idea is to minimise the negative consequences. These experiments should be done under specific conditions and with specific guidelines in mind. Like the problem we saw with Lindblom, such experiments might be too narrow, and therefore they can miss the bigger picture.

Lastly, *mediation analysis*, is about the study of the change of values itself. Kudina and Verbeek [87] mention the way that our perception and understanding of privacy might change through the addition of augmented reality in for example smart glasses. These artefacts create what is real to us [104], rather than convey reality. The mediation itself is in some sense an ontological status which provides a perspective on the world. The question of embedding is, in this regard, a debate about morality in the making, considering that we want these technologies to work for the greater good, we want to make sure that the impact they have on our perspective of society leads to the betterment thereof. As Kudina and Verbeek apply the theory on technomoral change, they show what the impact of technology may be on morality. It is to lay bare which perspective is acceptable and which is not. The problem therein lies that certain effects may only expose themselves once they have been thoroughly embed-

ded on a large scale in society, and that may not immediately arise from the technology itself but rather from the off-shoots thereof. Even if certain values are changed through artefacts mediating the world, the complexity of our societal systems may prove incomprehensible to project onto what a subtle mixture of artefacts may cause.

2.2.3. THE TECHNOLOGICAL EMBEDDING OF TECHNOLOGY

What we can glean from these different debates is that they run into very similar issues. Anticipation is like trying to solve path-dependency upfront, it contains an amount of speculation which indeed could blindside us. While updating during implementation is like doing small localised sociotechnical experiments, which may not sufficiently consider that the entire experiment itself may not be worthwhile to pursue. Mediation analysis may not give sufficient scope either, but it does investigate the changes during their happening. It is at the early adoption phase, which like sociotechnical experiments may overlook the problems which only arrive at the scene on a large-scale implementation. For example, think about the consequence of motor vehicles on the climate, their CO2 emissions are a problem in large fact due to their widespread use. Or how the internet was intended for the spreading of information, but also lead to misinformation, phishing, cyberbullying. This is obvious with the power of hindsight [105].

In short, solutions which drive to steer technomoral change suffer from the same issues that we see in path-dependency. The issue with oversight and corrigibility in both path-dependency and technomoral change is the fact that, down the line (built on other technologies), there may still be technologies that create an unwanted effect. It may be nice that we have invented the camera, and this may have a positive effect on society, but if it leads to the mobile phone with a camera in it and the coinciding sousveillance [106] (which has both a technological and a philosophical component in terms of privacy), then we may also need to consider to which extent the camera itself was worth the problems it created down the line. What is similar to both debates is that we need to draw a line somewhere in the sand at which we say things are beyond the scope of our purview. In such a regard, the anticipation of potential effects may be simply too speculative. Yet, mediation and sociotechnical experiments, may remain short-sighted. Similarly, one cannot solve the issue of path-dependency effectively by figuring everything out ahead of time nor by trying to evolve afterwards.

In the optimal case, we desire a change that is both good in the short term and helps towards a long-term solution. If we consider the issue of path-dependency as a search space, it may be that the local solution (short iterative cycles) may lead to a bad local optimum which is terrible to get away from. Whereas a heuristic from something like anticipation may lead nowhere worthwhile. For example, the invention of antibiotics killed the interest

in research for bacteriophages as a solution to illnesses [107]. Now we see drug resistant bugs in hospitals. Was it a worthwhile invention? This is highly dependent on the values that we use to view it. Antibiotics have saved countless lives, no doubt. Yet, they are also indicative in spawning a new kind of bacteria that we would have been better without. The question remains, would the research in bacteriophages have been any better?

What ought to be clear is that the question of technological embedding does not stop once technology is implemented. For one, with path-dependency in mind, it becomes clear that we may create new technology which subsists on previous generations. Yet, that doesn't mean the originator ought to become an irremovable object. The Collingridge Dilemma, treats it, in part through entrenchment, as mostly immobile once technology has been generally embedded. In essence, with path-dependency it is an ever-increasing kind of embedding. With more and more dependencies, it becomes ever harder to remove. Consider what it would take to remove all paper and paper-dependencies and paper-adjacent technologies (writing, but also computers with word processors, the printing press, libraries, and so on). Is it proper to consider the problem this way? Does that not strip us of our power to conceive of different alternative societies?

Perhaps the most poignant part of this is that technomoral change and entrenchment imply that we may sometimes be losing our perception of things because of said implementation of technology. If our perspective of both the world and ourselves is constituted by the technology, then we may endanger certain paths of exploration not only in a technological sense but also in a societal and ethical way. The technology may hamper how we can perceive alternative futures, both based on that technology but also in a larger perspective. Especially if we consider them as active mediators [104, 108]. The point of mediation in this case is not only that our perception of the world is altered, so is our ability to conceive of different worlds. In this ability to conceive alternatives, path-dependency and technomoral change meet. We may not be able to conceive of some alternative future because we have become stuck in the current reality with our current conceptualization to boot. This is precisely why I mentioned the extinction of experience [88] in the introduction, as it relates to this quite heavily. It is the feedback loop of interactions with mediators who play a shaping role in our perception of what is available to us.

This line of thought is found in Heidegger's [109] ready-to-handedness, meaning that technologies draw us in to certain kinds of actions, which are directly related to technological affordances [55]. It is also related to Latour [77, 110] whose conception of technology in Actor Network Theory implies something about the action of actors and how they shape relations. It is also found in Idhe [108], whose mediation of perception and technology is directly related to Verbeek's view [111]. Yet, what is lacking about the longer-term view? How can we make sure we understand the morality of technology, without en-

ding up implicitly accepting the framework it provides? The entrenchment happens even if we somehow remove the technology from society because *we* as a society are altered by it.

Before moving onto solutions, we need to discuss how AI plays into this . I have drawn broad strokes about technology, but AI plays an interesting part in the acceleration and change of this premise as well.

2.2.4. WHAT ABOUT AI?

Thus far, I have talked about general technological devices and systems. So what sets AI apart? For one, it takes far less time to converge and become entrenched. Consider, for example, the speed at which ChatGPT changed the work floor. We need to review how quickly such alterations converge. This is not the implementation of an infrastructure which takes decades or years to build but rather a year or months.

Another part of AI is that it exacerbates the self-reinforcing effects. There are temporal biases, namely that there is a preference for what currently exists in the data, if this feeds into a decision-mechanism, then it may become a prescription instead (accidentally or otherwise). Yet, the way that AI systems play an active part in deciding the content we arrive at (e.g. social media algorithms, google (scholar) search), may also invariable alter our perception as certain kinds of content are consistently favoured over others. In short, the self-reinforcing effect is worse because the consistency of AI systems greatly enhances some of the already present problems in path-dependency and technomoral change.

The way we make use of AI seems to accelerate embedding and also makes it harder to conceive of alternatives. Such systems tend to be highly complex and lack the opacity to actually understand how our perceptions may be changed in meaningful ways.

Even when we approach this through purported solutions, how far should we look ahead? What should we take into account? What do we consider good change, and under which conditions? And is it acceptable if such artefacts create the breeding ground for terrible ones? Furthermore, how should we even perceive the notion of alternatives if our moral framework is altered by the availability of said artefacts? In both debates, we run into the fact that we are limited in knowledge, time and rationality. So, the question is, what can we do about it?

2.3. DESIGN FOR PREFERABLE ALTERNATIVES

If we desire to design better technology that takes the long-term harms into account, then we need to ask who is designing. One could argue that the things we can consider, which factors matter, and how the problem is devised as well as the capacity to speculate about the future and long-term consequences, are at the very least dependent on the designer who is studying the

issue.

So, one way the question can be posed is: who gets to design what becomes entrenched? And do they accurately consider the consequences that follow in the long run? If we just assume that any designer and design is worthwhile, then we create a kind of "anything goes" attitude. At its worst, this results in market-driven societies where one may prioritise profitability over ethical considerations, no matter the implications. These two points illustrate just a short-term view, which may still ignore the larger issues entirely. Furthermore, the degree of openness present in such an attitude does not align with the ideas which try to alleviate the Collingridge Dilemma regarding technomoral change. We do not merely want to study how values change through technology, but also steer them in the right direction [63]. Mediation analysis, sociotechnical experiments, and technomoral scenario writing are all still in good part delineated by the ones who are doing the designing. For path-dependency, it is much the same. We see that we need to wrest free from path-dependency such that we do not get locked in, but it does not mean we know where to go. Again, when taken together with the notion that technological devices can be considered constitutive of what it means to be a human [112], we can understand that this entails a lack over our own perceptions and moral framework.

It is helpful to understand the choice involved in what becomes entrenched and why. Take, for example, the effects of the industrial revolution on art. The method of making oil paints and paper once belonged to the secrets of an artist's studio, but with the invention of paint tubes and mechanised production of paper, we saw two things: it became widely available and the capacity to create it by hand significantly dropped among artists [103]. Furthermore, because quality had rapidly deteriorated too during the production, the capacity to conserve art also severely diminished. The effect was the widespread availability of paper and paint, which perhaps a century before no one would have deemed worthy of use. In essence, cheap and available paper could have uplifted commoners to allow them to paint, a kind of worthwhile equality for all. Yet, it came at the sheer cost of quality. No matter if it were arbitrary or due to the ideal of profit maximization, there was a preference of availability and price over quality. This is not a moral argument per se, but it is rather an open question about the nature of man and its society. The anything goes mentality does not consign itself to anything of morality. In terms of technomoral change, it means that any change can be considered as good as the next. Since we may not know what good or evil, it may bring down the line.

Another example may be the development of ChatGPT. It can indeed increase the ease of our work with text, but it comes at the cost of deskilling certain types of work (e.g. writing and summarizing), It can also be a headache for teachers supervising students who use such technology and who want to test them on such skills. Another is the means of transportation. Travel-

ling great distances was made easy by certain types of transport like the train and planes, which came at the cost of overcoming the difficulties that travel used to entail. The lightening of labour necessarily comes at the cost of the skill required to do that labour before. To paraphrase Thoreau: it may be nice that we can talk to anyone in the world, but we need to have something to discus [113]. In terms of technology, it appears that it may be nice to have efficiency and effectiveness and capacity, but it needs to be subservient to a worthwhile goal. If we take that to be something along the lines of human flourishing, then we need to consider that overcoming difficulties and maintaining quality is a good part of human flourishing too. If we desire to avoid such moral relativism, then we need a framework of which technology is morally permissible and which is not. In essence, we need a system on which to decide on what basis we should accept technology.

While this may seemingly sound overarching in a way that requires an uncanny understanding of *the good life*, we do not propose to solve that issue in such a way. If we do accept all the clauses of the Collingridge Dilemma, then we must also admit that there is no way we can truly oversee whether entrenchment may lead to terrible technology down the line. Yet, the fact that entrenchment is unwieldy is something we can partially dissipate. However, there are ways we can deter the problems of entrenchment rather than address entrenchment itself directly. A *design for alternatives*, would allow some of the functionality to be standardised in such a way that it can be more easily replaced, even if other technologies are dependent on it. Designing for alternatives requires us to think about questions regarding the *interoperability* of these alternatives, the *scalability* of the implementation of them, and the mediative responsibility of designers, which impose such changes on our perception.

2.3.1. INTEROPERABILITY

One way to think about this is the difference between type and instances[2]. E-mail provider for example is a type, of which ProtonMail, Gmail, and many others are instances. E-mail makes use of an open standard (e.g., SMTP, IMAP, POP) such that all these different instances can send messages towards another. This allows an e-mail client to be replaced if it proves unwelcome. The beauty of open standards is that the technology that ensues from it is interoperable by design [114]. This solves the entrenchment issue for instances, if done well, since the barrier to swap from one client to another is significantly lower if one can still mail everyone else. If we offset this against messaging apps (Currently Signal does not talk to Whatsapp), then we understand how the cost of changing between service providers, namely the loss of one's con-

[2]The basic premise of type and instance, is that a type (within computer science) is the general category, say a car, whereas an instance is a particular entity of a type (e.g. a specific BWM)

tacts, may invariably change depending on the use of an open standard[3].

2.3.2. SCALABILITY

Yet, the problem of entrenchment also arises within a type. For example, if we found out that e-mail had a disastrous consequence on our feeling of connection, then, even if we could replace all the e-mail clients in the world, the problem would remain. It would be e-mailing itself that would cause the issue. In such a case, interoperability may provide leeway to alleviate some of the worries (for example, introducing additional restrictions). Yet, if that proves insufficient, then we need to understand that we cannot press an "undo button" and have never invented e-mail. Designing for alternatives works best, once we have the right scalability of such types in mind. If certain issues arise from large-scale implementation, then perhaps we should avoid implementing them on such a scale. Because if an open standard or a type becomes an oppressive and dominant feature, and we are unable to deviate from it, then the thing itself as an entrenched entity may still cause issue as newer technologies are developed.

The need for interoperability and limited scalability mitigates the potential harm a designer may bring to the table. If the application is wrong, it can be swapped out and if the open standard turned out to be terrible, then the harm is limited to a particular bandwidth. Of course, limited scalable is no foolproof solution to the problem of entrenchment as certain things we would like to have on a larger scale, nor does it answer what the right scale of certain standards would be (even though that is likely dependent on context). Furthermore, certain problematic effects may be present but only become apparent once it is one a large scale. The risk is that the trouble may already exist on a small scale, but our chances of perceiving it could be hindered.

Drawing from the previous section, we can compare interoperability and limited scalability to the solutions proffered to solve path-dependency. Interoperability is an upfront idea, namely an open standard, which, if also not fully implemented on a large scale, could even be iterated upon (during the process). Overall, if something is wrong, both limits of scalability and interoperability allow for the regaining of scope afterwards. There is thus a measure of upfront consideration, a potentially iterative cycle in the middle, with additional restrictions afterwards to overcome issues caused.

2.3.3. MEDIATIVE RESPONSIBILITY

The solutions of open standards and interoperability, say nothing about whether we can revert our perception of Good and Right (even though moral truths may be invariant, our perception thereof is certainly not). Innovations

[3]It should be noted that ecosystems, like the interconnectedness of Gmail and Google calendar undo much of the benefits of an open standard as users still get locked-in by use of such an ecosystem.

interact with other commitments, technologies, and moral perceptions, and it is an impossibility and an implausibility to counteract this, as we are altered by a plethora of things. It is similar to the way a text can alter our perception of the context we are in. The horizon of what we can perceive and understand is in constant flux[4]. Yet, this means that even technologies that hold to interoperability and limited scalability may still alter our perceptions to an extent that such a change becomes entrenched anyway in our manner of thinking.

Because our environment is altered both directly and indirectly through countless innovations, it means we need to ask what is changed exactly and by whom. Yet, the real issue comes in when we lose the ability to understand that there are conceivable alternatives that are worthwhile to investigate and discuss, much in the way that Fischer describes in Capitalist Realism [116]. In other words, it may be possible that we lose the capacity to ask such questions[5]. It is fundamentally this part of entrenchment which is so unwieldy in the Collingridge Dilemma. It implies it is far too difficult to understand how many different conceivable worlds there are, and it is exactly this for which we need visionaries who can show us that there are possible different alternatives, with and without particular artefacts. We need clearer channels for dissent, that allow for true destruction and alteration of technology.

Furthermore, essential to any kind of technological innovation and implementation is thus showing that the perceptions created are not set in stone. This is much akin to what Verbeek calls the opacity of context [104]. Mediation theory belies the argument that implementation of technology, because of its alteration of our moral perceptions, is applied ethics [111]. If we accept this, then we require an answer to sheer moral relativism.

Another route is responsibility of designers. It is an often heard argument that use of technology is not the responsibility of designers, since we often hear that we cannot reasonably hold designers responsible for the use of a piece of technology. Yet, given the fact that technology elicits both a certain kind of response and action and an alteration in perception, we *should* perhaps attribute responsibility of such elicitations to developers. For one, this may lead to something like Achterhuis's conception of technology [117]. In which technology itself is moralised (e.g. it should act in a particular way). The problem with this, even though it may be true, is that it can be seen as an attack on human freedom [111]. Rather than immediately suggesting how technology ought to look, we can hold designers responsible for their mediative consequences.

Given the Collingridge Dilemma, we may not be able to have complete

[4]This is meant as a reference to the hermeneutic circle [115], though to discuss both hermeneutical technological assessment and Gadamer, either is far beyond the scope of this paper.

[5]One clear hallmark of the lack of conceivable alternatives also falls within the lines of our discussion. Technological determinism, namely the entire idea that technological innovation is inevitable, is precisely the kind of thinking which does not allow for alternative worlds to be conceived of easily.

control over how the technology develops and is used, but we can certainly hold developers responsible for the design they unleash and the consequences they entail. Perhaps this requires a kind of timespan for which the designer can be held responsible for any complications that arises (e.g. something in the same vein as copyright protection). This may cause designers to be hesitant to wildly innovate for the wrong reasons. Instead, it may promote designers to actually live up to the ideals of designing with values in mind.

2.3.4. GOVERNANCE

Interoperability, limited scalability and mediative responsibility all fit with a notion of Regulation by Design (RBD) [118]. The idea of RBD is that we also have to entertain the possibility that regulation is an active participant in shaping behaviour. Design comes into play with this, as it can be a critical component of regulation, providing another modality with constraints and affordances [119]. The application of interoperability, scalability and mediative responsibility begs the question of how this relates to governance. Can this be applied? What is an acceptable scope? Thus far, this has been an appeal for design, but can this be enforced? In some sense, yes. What this requires is a reconceptualization of the idea of a monopoly. Design is inherently monopolistic. It's enforcement of affordances and constraints is by its very nature not open to competition. What we need is design that has a better possibility of being replaced, and for that, we can enforce open protocols, standards and, for example, argue that there is a maximum number of users a platform can have.

2.4. CONCLUSION

In this chapter, we spoke about entrenchment through the lens of path-dependency and technomoral change and saw that they had relatively similar issues and solutions. The problem remains, what we do with entrenchment as dependencies arise. As our moral perception is altered, the coinciding alteration is harder and harder to reverse and becomes more incoherent to think about. We discussed how the various solutions together may lead to an idea of interoperability and limited scalability, but these do not deal with the fundamental issue of self-reinforcing effects completely. In short, if we want to deal with the fallout of path-dependency we need to be able to conceive of different worlds with different technological devices, but it is exactly this which is harder to do when innovations have become embedded. To alter this, we need better conceptions of responsibility and better channels for dissent.

BIBLIOGRAFIE

[11] D. Collingridge, “The social control of technology”, 1982.

[39] P. A. David, “Path dependence, its critics, and the quest for ‘historical economics’”, in *Evolution and path dependence in economic ideas*, Edward Elgar Publishing, 2001, p. 15–40.

[55] W. W. Gaver, “Technology affordances”, in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1991, p. 79–84.

[63] T. Swierstra, D. Stemerding en M. Boenink, “Exploring techno-moral change: The case of the obesitypill”, *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments.*, p. 119–138, 2009.

[76] S. Zuboff, “The age of surveillance capitalism”, in *Social Theory Re-Wired*, Routledge, 2023, p. 203–213.

[77] B. Latour, *Aramis, or the Love of Technology*. Harvard University Press, 1996.

[81] J. Danaher en H. S. Sætra, “Mechanisms of techno-moral change: A taxonomy and overview”, *Ethical theory and moral practice*, jrg. 26, nr. 5, p. 763–784, 2023.

[82] J. Newman, M. Mintrom en D. O’Neill, “Digital technologies, artificial intelligence, and bureaucratic transformation”, *Futures*, jrg. 136, p. 102 886, 2022.

[83] B. Alcott, “Jevons’ paradox”, *Ecological economics*, jrg. 54, nr. 1, p. 9–21, 2005.

[84] P. Garratt, “Sublime transport: Ruskin, travel and the art of speed”, *Travel Writing, Visual Culture and Form, 1760–1900*, p. 194–212, 2016.

[85] J. Stilgoe, R. Owen en P. Macnaghten, “Developing a framework for responsible innovation”, in *The Ethics of Nanotechnology, Geoengineering, and Clean Energy*, Routledge, 2020, p. 347–359.

[86] N. Pleasants, “The structure of moral revolutions”, *Social Theory and Practice*, p. 567–592, 2018.

[87] O. Kudina en P.-P. Verbeek, “Ethics from within: Google glass, the collingridge dilemma, and the mediated value of privacy”, *Science, Technology, & Human Values*, jrg. 44, nr. 2, p. 291–314, 2019.

[88] M. Soga en K. J. Gaston, "Extinction of experience: The loss of human–nature interactions", *Frontiers in Ecology and the Environment*, jrg. 14, nr. 2, p. 94–101, 2016.

[89] K. J. Gaston en M. Soga, "Extinction of experience: The need to be more specific", *People and Nature*, jrg. 2, nr. 3, p. 575–581, 2020.

[90] A. Genus en A. Stirling, "Collingridge and the dilemma of control: Towards responsible and accountable innovation", *Research policy*, jrg. 47, nr. 1, p. 61–69, 2018.

[91] W. B. Arthur, "Competing technologies, increasing returns, and lock-in by historical events", *The economic journal*, jrg. 99, nr. 394, p. 116–131, 1989.

[92] G. Schreygg en J. Sydow, *The hidden dynamics of path dependence: Institutions and organizations*. Springer, 2009.

[93] J. Sydow, G. Schreyogg en J. Koch, "On the theory of organizational path dependence: Clarifications, replies to objections, and extensions", *Academy of Management Review*, jrg. 45, nr. 4, p. 717–734, 2020.

[94] T. Swierstra, "Nanotechnology and technomoral change", 2013.

[95] W. H. Sewell, "Three temporalities: Toward an eventful sociology", *The historic turn in the human sciences*, jrg. 98, p. 245–280, 1996.

[96] K. Levin, B. Cashore, S. Bernstein en G. Auld, "Playing it forward: Path dependency, progressive incrementalism, and the "super wicked" problem of global climate change", in *IOP Conference Series. Earth and Environmental Science*, IOP Publishing, deel 6, 2009.

[97] C. E. Lindblom, "The science of "muddling through"", *Public administration review*, p. 79–88, 1959.

[98] J. Fortwengel en A. Keller, "Agency in the face of path dependence: How organizations can regain scope for maneuver", *Business Research*, jrg. 13, p. 1169–1201, 2020.

[99] A. Grunwald, "Technology assessment: Concepts and methods", in *Philosophy of technology and engineering sciences*, Elsevier, 2009, p. 1103–1146.

[100] I. Van de Poel, "Why new technologies should be conceived as social experiments", *Ethics, Policy & Environment*, jrg. 16, nr. 3, p. 352–355, 2013.

[101] ——, "Nuclear energy as a social experiment", *Ethics, Policy & Environment*, jrg. 14, nr. 3, p. 285–290, 2011.

[102] ——, "An ethical framework for evaluating experimental technology", *Science and engineering ethics*, jrg. 22, nr. 3, p. 667–686, 2016.

[103] W. Kemp, *The desire of my eyes: The life and work of John Ruskin*. Macmillan, 1990.

[104] P.-P. Verbeek, "Expanding mediation theory", *Foundations of science*, jrg. 17, nr. 4, p. 391–395, 2012.

[105] R. Cecchi, T. M. Haja, F. Calabro, I. Fasterholdt en B. S. Rasmussen, "Artificial intelligence in healthcare: Why not apply the medico-legal method starting with the collingridge dilemma?", *International journal of legal medicine*, p. 1–6, 2024.

[106] S. Mann, J. Nolan en B. Wellman, "Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments.", *Surveillance & society*, jrg. 1, nr. 3, p. 331–355, 2003.

[107] W. C. Summers, "The strange history of phage therapy", *Bacteriophage*, jrg. 2, nr. 2, p. 130–133, 2012.

[108] D. Ihde, *Technics and praxis: A philosophy of technology*. Springer Science & Business Media, 2012, deel 24.

[109] M. Heidegger, *Being and time*. SUNY press, 2010.

[110] B. Latour, *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford, 2007.

[111] P.-P. Verbeek, "Materializing morality: Design ethics and technological mediation", *Science, Technology, & Human Values*, jrg. 31, nr. 3, p. 361–380, 2006.

[112] A. H. Kiran, *The primacy of action: Technological co-constitution of practical space*. Norges teknisk-naturvitenskapelige universitet, Det humanistiske fakultet . . ., 2009.

[113] H. D. Thoreau, *Walden*. Yale University Press, 2006.

[114] M. Sjarov, D. Kißkalt, T. Lechler, A. Selmaier en J. Franke, "Towards "design for interoperability" in the context of systems engineering", *Procedia CIRP*, jrg. 96, p. 145–150, 2021.

[115] H.-G. Gadamer, *Truth and method*. A&C Black, 2013.

[116] M. Fisher, *Capitalist realism: Is there no alternative?* John Hunt Publishing, 2022.

[117] H. Achterhuis, "De moralisering van apparaten", *Socialisme en democratie*, jrg. 52, nr. 1, p. 9–18, 1995.

[118] K. Prifti, J. Morley, C. Novelli en L. Floridi, "Regulation by design: Features, practices, limitations, and governance implications", *Minds and Machines*, jrg. 34, nr. 2, p. 1–23, 2024.

[119] L. Lessig, "The new chicago school", *The Journal of Legal Studies*, jrg. 27, nr. S2, p. 661–691, 1998.

3

How to Gain Control and Influence Algorithms: Contesting AI to Find Relevant Reasons

One day I was watching a paramecium and I saw something that was not described in the books I got in school - in college even. These books will always simplify things so the world will be more like they want it to be.

Richard Feynman, *Surely you are joking, mr Feynman*, p.81

Relevancy is a prevalent term in value alignment. We need to keep track of the relevant moral reasons, we need to embed the relevant values, or we need to learn from the relevant behaviour. What relevancy entails in particular cases, however, is often ill-defined. It is hard to define relevancy in a way that is both general and concrete enough to give direction towards a specific implementation. In this chapter, I describe the inherent difficulty that comes along with defining what is relevant to a particular situation. Simply due to design and the way an AI system functions, we need to state or learn particular goals and circumstances under which that goal is completed. However, because of both the changing nature of the world, misalignment occurs, especially after a longer amount of time. I propose a way to counteract this by putting contestability front and centre throughout the lifecycle of an AI system, as it can provide insight into what is actually relevant.

3.1. INTRODUCTION

Artificial intelligence (AI) systems should be aligned towards societal good. Nonetheless, mistakes are racking up, and so there are attempts to get a better grip on how to implement and control AI systems. The result of this search for better AI systems and control over them is not without merit, as it has led to numerous theories about how one ought to control and/or design such systems [16, 22, 23, 100]. There are also good reasons to search for a solution to misalignment, as such systems cause serious harm [65, 120, 121] or prove detrimental to institutions [122].

Much of the troubles designers have with their implementation has to do with moral reasons. Designers need to weigh or trade-off certain values, understand a treasure trove of contextual information, and discern between a set of minute details that may not at all be clear at first glance. The saying: the devil is in the details, could not be more true with regard to value-alignment. At the end of the day, it is not the high and mighty ideals of the designer that matter, but the down on the ground implementation affecting people's lives, and when there is serious harm, then these may also inadvertently skew the public debate such that AI systems can become unwanted.

What I aim for, in this chapter, is to alleviate some of the issues involved with implementing value aligned strategies. As I shall show, in practice it may be difficult to determine what reasons are relevant and which are not. This is not only due to contextual factors, but also who is involved in the design process. I propose to include contestation at different stages of the system design to improve the situation and application of AI systems, and I offer a mostly agnostic way to adjust the system based on missing inferences.

In short, I will first go over some difficulties with the implementation of value alignment. What we will see is that it requires knowing relevant moral reasons. Yet, these are difficult to find because relevancy is not a given. If we assume that to be the case, then we implicitly make assumptions about what is relevant and what is not, likely leading to unaligned implementation. The issues with relevancy are two-fold: (1) Theoretically, formalization has certain issues and (2) designers are limited in knowing what is relevant, even if they talk to stakeholders. To counteract this, I introduce contestation throughout an AI lifecycle to better align such systems, but this requires that contestation leads to meaningful adaptation.

3.2. WHEN WE TALK ABOUT RELEVANCY

Value alignment may be difficult to implement because of relevancy. We need to find the relevant moral reasons, but if a designer does not know those or understand them and has no means of finding them, they will be hard-pressed

This chapter was originally published as: Kuilman, Sietze Kai, et al. "How to gain control and influence algorithms: contesting AI to find relevant reasons."AI and Ethics 5.2 (2025): 1571-1581.

to ever create a value aligned system. Relevancy is, in that regard, the bedrock of most value aligned theories, as they deal with the matter of correctness. It is not always transparent what and when something is relevant, so without any good substantiation of such a notion we can provide beautiful theories with little to no effective application, as we may not have the means to find the correct moral reasons for a given situation.

Of course, we need to first understand what we mean by moral reasons. The nature of a moral reason is dependent on one's views of morality. One can view them as a determinant which plays a part in one's action. For example, lying is wrong, gives us a duty not to lie, but the reason why we ought not lie could be: never treat another as a means to an end. This is of course a very different presupposition compared to: honesty maximizes happiness. Both leading to the behaviour of honesty, but in particular situations these underlying beliefs matter for how a problem is approached and what particulars are used in a solution. Moral reasons are thus about the domain of correct action (e.g. prevention of harm). Normativity is a broader in scope, which could be addressed here as well, but I dive in particular into the notion of what one ought to do.

A central topic of moral reasoning is also casuistry, the search of finding the relevant moral conditions and discerning the more relevant ones from the only slightly relevant ones. If a designer is incapable of making relevant distinctions in context or understand what needs to be kept track of in said context, then it doesn't make a difference if they try. Furthermore, if those relevant distinctions are not presentable in such a system (e.g. those features cannot be captured because of their complexity or inherent indescribability), it will most likely cause the same problem. The question for correct behaviour is thus:

> *What relevant moral reasons do we require for such AI systems to function properly?*

For computer science, relevancy has been of much importance within information retrieval, as finding the correct and relevant document has been vital to the field [123]. If the relevancy of a document is determined by the number of clicks it gathers (say on the website of a search engine), then we can safely say that older documents which have been exposed to time are likely to have generated more clicks, in that regard an underlying belief is that older documents may be more relevant. Certainly, we can understand this for literature, where classics have to withstand the test of time, but this may not be appropriate for certain types of documents (say scientific information). The quest for a designer is to know which reasons are relevant and how they should be implemented such that the right documents fall into the right hands.

In this chapter, I will take one route to relevancy, but I believe there are many more possibilities. However, the main premise I postulate in terms of

relevancy is that during design we need to fill in the details. We make choices when applying a (moral) theory about what is relevant. I argue that, without an adequate idea and application of relevancy, we are at a loss of finding effective applications as the openness in implementation ends up detailing the most important part.

3.2.1. VALUE ALIGNMENT

Theories on value alignment are built upon the proposition that AI should not merely act, but should also act such that certain harms can be avoided [15] or that they positively influence a kind of human flourishing [124–126]. How do we make sure that machines act in accordance with our values? There are two basic approaches to this: sociotechnical solutions [16, 22, 65, 127, 128] and more technical ones [23, 129–131]. Sociotechnical solutions involve users and try to picture the machinery such that there can be an interplay between humans and artificially intelligent agents. Technical solutions stem from the belief that value-aligned action can be seen as a technical problem and can be solved as such. While there are obvious differences between sociotechnical solutions and technical ones, the main premise of this chapter holds for both approaches. The particulars of technical solutions, still requires data, features, and some line by which to draw what is relevant and what is not. The inference of a pattern from said data and features also requires much more understanding in concrete cases than simple assuming that this relevancy is easily found or induced appropriately.

Applying value alignment theories in any domain is a feat dependent on context, goal, and structure. For example, knowing whether an application is discriminatory requires that we also know whether that discrimination is at any point acceptable in that context, meaning that we need to disambiguate discrimination as discernment (to delineate different options) and discrimination as unjust bias (to categorize groups on features that are deemed inappropriate). We may want to discriminate (discern) between different groups, but we do not want to discriminate individuals (unjustly). This is essential because we may sometimes really want to recognize a sick patient from a healthy one. However, we don't want a system to only recognize sick patients of a particular gender (unless the disease is gender-specific, of course). Considering the context, we need to know whether we are actually introducing unjust bias or doing the right inference. This means we have to know at which juncture it is one or the other.

Such knowledge requires a particular kind of oversight and knowledge of the system. The problem of reward hacking in terms of goals [132] - that being, the AI system finds a misinterpretation of the goal such that it can maximize its reward function - presents a serious issue to this kind of knowledge, as an AI system may optimize for something (unintentionally) through unjust means. The context in which it is placed is also highly important. Facial recognition

is not necessarily unaligned, but if it is applied in a way to arrest a particular group, then we can talk of the relevancy of being able to discern that group and the misalignment of that in the face of human flourishing.

On top of that, designers do make choices about what is relevant and what is not, both on a technical level and on a social one. If a designer wants a recommender system to be value aligned, then they need to know what it ought to be recommending and what they can recommend. This entails a kind of idea about what the most relevant detail may be within possible documents and how to extract it. Such a detail needs to be discerned from the context, meaning that the context actually does need to contain that detail. If the data is structured in a way that does not allow correct or full access to the relevant detail—or does it in a way that ties it to other factors, then they are bound to infer a different pattern than the actual relevant one. A choice in recommending based on citations or recommending based on number of views, is likely going to result in very different recommendations. While we can debate whether that is moral perse, this is easily transposed towards a moral domain by changing clicks and cites into recommendation based on sensitive topics. For example, demoting climate scepticism or certain information about war.

Not considering these issues is not a way out. To assume the data collected is correctly structured and always contains the correct scope of information, even humanly labelled data, to reach the correct goal, is quite a set of hefty assumptions to make. It does not mean value alignment becomes easier, in fact, it only means that one has made implicit assumptions about what is relevant on a technical level and on a social level, without delving into it.

3.2.2. RELEVANCY

Having argued for the necessity of relevancy, we need to understand what we need to know about relevancy. However, this is somewhat difficult to define. When is a reason relevant?

Relevancy is found in a plethora of fields and studies, such as: communication [133], logics [134, 135], and information retrieval [123, 136]. It will most likely play a part in many goal-oriented studies. Finding the correct treatment means knowing certain relevant facts about a disease. Knowing how to construct policy means knowing (some of) the relevant actors in a particular case. To find the relevant moral reason requires knowing what they are or how to find and evaluate them. The point of constructing or excavating a concept like relevancy in value alignment is not a quest to survey all potential relevancies. Rather, we desire to know what kind of relevancy we could look to.

As we mentioned in the section on value alignment, there are contextual features, saliency, and teleology to keep in mind when discussing relevancy. How we use an AI system is of importance to value alignment. And while we could discuss the specifics and lack of oversight of applications as a serious

issue, for this chapter we will assume this is about intended use. Where it is used and for what purpose it is used, seem like obvious parts to relevancy. This, however, does not explicitly cover saliency. Yet, if a relevant distinction is not salient to the system, then it will likely not be able to draw the right inferences from the context.

AI systems hold a particular position within societal, governmental, or commercial institutions which is very different to that of humans and this needs to be kept in mind. As the relevancy of human beings may change and update on the fly, while that of a system (even online ones) have been trained in a particular fashion and have a certain depth and breadth of possible implementations. During design, one needs to scope what the possible actions for an AI system are, what its range of contextual features are, and what is salient to it. In the broadest sense, in value alignment, we seem to be asking for the impossible. As Turing already noticed[137]:

> It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty may well arise from this decision later. To attempt to provide rules of conduct to cover every eventuality, even those arising from traffic lights, appears to be impossible.
>
> - Alan Turing, *Computer machinery and intelligence*, p.16

One should note that Turing's comment also works for learning relevancy in machine learning models. The decision of data and the model used within a set of circumstances may be inappropriate in another context. Such limitations make it impossible to produce the induction of a function (rather than a set of rules) which allows the machine to act in every conceivable set. [1]

In the ideal setting, we would be able to cover every eventuality, then our systems would be completely value aligned, acting appropriately in all edge cases and novel circumstances. Yet, we also need to see what relevancy means if we are pragmatic. To reiterate on the example of recommender systems from the previous section, if we desire that recommender systems do not spread misinformation, we need to evaluate what content they recommend,

[1] A similar distinction of conceivability is also made in Philosophical Investigations [138], §193, in which Wittgenstein discusses the difference between a machine as a symbol and a machine in terms of its behaviour. Furthermore, there is evidence that Turing was also aware of Wittgenstein's position on some of these issues[139]. In short, the comment by Turing should not be seen as an attack on the method by which we arrive at the behaviour, rather it is an argument against the possibility of describing such behaviour at all.

and this shifts with time. The scientific community may gather a new view on things or a definition of misinformation may miss a particular new piece of misinformation due to its novelty. If we desire to value integrity and honesty (as values which we take to be at least some basis of why we want to avoid giving misinformation) then this boils down to defining what those values mean (not sharing misinformation). Yet, that is simply not a static conception nor a static output. Things that are misinformation may turn out worthwhile and vice versa. When something is relevant has thus much to do with its context and its intended use. A proper functioning of an AI system thus means that we can detail what exactly we need from it, in terms of delineating context and inference patterns in relation to its goal.

Yet as noted, relevancy of these systems comes in a particular way and this opens us up to two problems: the formalization aspect of such systems and those who are involved in the process of formalization. After we have delved into that topic in particular, we will spend the remainder of the chapter explaining what we can do to find the relevant moral reasons.

3.3. ADDRESSING THE ALGORITHM IN THE ROOM

The main reasons why we distinguish between relevancy humans display and those that belong to artefacts result from the formalization of context and use, and the fact that these artefacts are being designed by designers. These factors lead to issues that make finding the relevant reasons far from easy. We will describe why this may be the case in this section.

3.3.1. THE FRAME PROBLEM

In the sixties and seventies, a roughly similar problem as the one we describe with value alignment was addressed in terms of expert systems with logic statements. McCarthy and Hayes[140] recognized that there was a problem with representationalism, namely: there is a lack of inertia when dealing with predicates. Each time an update function was performed (to see if any predicates had changed based on action) all predicates had to be checked because there was simply no knowing which had to be updated and which had not. To think of this in simple terms, I can paint an object, but if I move it, how do I know it hasn't changed in colour? Going over each and every predicate was simply a waste of both calculations and space, because many things would not change given a simple action. Yet, not going over such dependencies might cause the machine to overlook simple yet important dependencies when predicates did change. They called this double-bind the frame problem.

The concept was quickly appropriated by a broader philosophical community, wherein the discussion was not specifically meant to address problems in logical calculus, but rather to address the question of relevancy and action [141–143]: How to act and update beliefs about the world? The "whole puddingöf the frame problem - meaning both the version McCarthy

and Hayes defined and the ones philosophers aligned with it - shows the practical limits of describing actions in terms of relevancy. The frame problem applies today, even in machine-learning, as we can ask how we ought to define the world, and what limits we need to draw in featurizing such that the pattern we achieve is correct. This is not a trivial task, and it may be the reason why we resort to using a term like relevancy in value alignment. It may be too difficult to know which moral reasons we need to account for. To show this difficulty, we base ourselves on Daniel Dennett's example [141].

> One day its [R1] designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. ... There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called `PULLOUT (Wagon, Room, t)` would result in the battery being removed from the room.
>
> - Daniel Dennett, *Cognitive wheels: The frame problem of AI.*, p.11

To put this in brief terms: We have a robot $R1$ and a task. $R1$ has a set of permissible actions. Each of these actions can be learned or formulated through logic, but all are meant to complete or work towards the completion of task. Dennett [141] discusses $R1$ does not understand all the important relations:

> Straight away it [R1] acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon.
>
> - Daniel Dennett, *Cognitive wheels: The frame problem of AI.*, p.11

Although the robot had a task and a set of actions, it had missed the relationship between the bomb and the wagon. As the creators understood, $R1$ had missed the relevancy of the context. So aside from the set of possible actions, each action should also be placed in a context such that: an action is desirable in context such that said action actually aligns and contributes to the completion of the task. Here of course the heart of the problem arises as the creators create another robot $R1D1$, that also deduces the relevant context for a specific action given a task.

> They placed R1D1 in much the same predicament that R1 had succumbed to ... It had just finished deducing that pulling the wagon out of the room would not change the colour of the room's walls, and was embarking on a proof of the further implication

> that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon - when the bomb exploded.
>
> - Daniel Dennett, *Cognitive wheels: The frame problem of AI.*, p.11

Thus, we see the double-bind arise from the frame problem. We overshoot or undershoot in framing. In terms of relevancy, the machine overlooks certain factors, or it dies trying. What we can distil from this is the following:

The frame problem: How does a machine recognize the correct context and determines the relevant features in said context such that all actions result in or contribute to the accurate and correct completion of task?

This definition of the frame problem seems to revolve around relevancy. With this relation in our mind, we can see certain problems in AI in a different light. There are plenty of examples of AI systems making mistakes with moral implications, e.g. classifying the entrance to Auschwitz as being related to sport [144]. These problems may be addressed through some theory of value alignment, however, that does mean we need to sufficiently address this topic of relevancy.

The simplest way to understand the exact nature of this problem can be viewed through the perspective that Turing also proffered. The simple fact remains that such an AI system won't have been tested under literally all possible conditions, meaning there could and most likely will be mistakes. And those responsible may not be capable of overseeing what the consequences will be. If, however, we knew the relevant features and conditions, and could model those accurately, then knowing the trajectory of possible actions may be possible. It requires that we understand what context is involved, what the system highlights (puts emphasis on, gives importance to) and what the goal is specifically aligned to[2]. In this regard, such systems may still be brittle in novel circumstances, but if we apply them correctly and hem them in, then we should be able to use them effectively[3].

The frame problem is an apt example when we desire to show the practical limits of value alignment theories without some interpreted notion of relevancy. Theories of value alignment may be helpful to decrease the scope

[2]Looked at from this perspective one can argue that the frame problem also essentially poses that value alignment in its ideal case is improbable if not infeasible. As we ourselves are also limited and may not have the capacity to really derive, formalize, or process what is relevant to a given situation. This is of course an abstraction of what value alignment is about - making better machines, which we can surely do by at least trying to incorporate these ideas and thoughts.

[3]This notion of correct application is also considered in the idea of a Moral Operational Design Domain[65].

of possible implementations, but it will practically still resort to some notion of relevancy to detail the context, teleology, and saliency. While there may be technical limits to the feasibility of certain aspects of relevancy (e.g. value trade-offs, or incompatible emphases, or simply intractable contextual scope), we do need to have a coherent and stable practice to assemble a reasonable grounding as to why certain choices were made throughout an AI lifecycle. Otherwise, we run the risk of maintaining a kind of *anything goes* attitude.

3.3.2. RELEVANT TO WHOM?

By now, we should have a better understanding of the problem of value alignment. We need the right (moral) reasons for AI systems to function properly, but these are not given. Designers are likely not getting it right first time. This problem of course also counts for institutions and policies as well, yet we also know these can have potential benefits. Without taking away the work we do during implementation, we should start attaching more thought to the life cycle of a system once it is nestled in its context.

Yet there is one more major discrepancy in terms of relevancy which needs to be brought forth. These systems have to deal with the contention between different stakeholders, users, end-users, whom all may have different views which are incompatible perhaps on the level of context, perspective, or goals. These AI systems cannot easily entail to the monotheistic view of relevancy because it does not take into account the veritable jungle of opinions that stakeholders may have. What relevancy entails in value alignment is not merely a disambiguation of contexts and correctly specifying teleological aspects. Rather, it is about what a group of stakeholders thinks is relevant during the design rather than what is relevant. Not only could the stakeholders, the intended use of such a system, or the context in which it sits, change over time, during design we also see the problem of relevance in terms of whom to invite to the table. Have we invited the relevant stakeholders?

One particular example that is interesting to note is recommendation systems for children, as recommender systems are mostly targeted at adults[145], yet children are also using these systems. If children were stakeholders (or parents concerned for their kids), designers need to consider more than merely the wants of the user. Instead, they may also need to incorporate very different dimensions, such as: educational, and developmental.

Yet, through contestation leading to adaptation, we could mend mistakes made during design. If individuals were capable of contesting an outcome and then having a kind of deliberation through or with the system, then some form of adaptation could be achieved, which could result in a better alignment of the system with the user.

3.4. CONTESTABILITY AND CONTEXT

As mentioned in the introduction, I think contestability provides a good way of counteracting the problems of relevancy that I have thus far discussed. I also mentioned that - under the right circumstances - contestability could provide a better aligned system for users, while also giving practical insight to designers about how they ought to adapt their system while it is operating. To understand what it is I propose, I also must understand some part of contestability.

Literature on contestability is often focused on giving an inch of control to users when faced with automated decision-making [146]. For example, human intervention requires that one is able to contest the outcome of a decision before it is enforced. It can be seen as a kind of procedural justice, ensuring that participants have a voice in the matter. But human intervention may not be enough for some, it may also require that people can fully grasp the outcome, linking it directly to explainability [147]. If individuals are given an inkling as to why the outcome is what it is, they may be more substantive and understanding of what is going wrong during the decision-making [148]. In all, contestability focuses on the illegitimate or unjust decision that can arise from automated decision-making.

While this is certainly a good point to make, and in terms of user empowerment it is an interesting tool, there remains an open question as to what designers should do specifically after contestation has arisen. How should the system be changed or updated, and how should that be done? It is a question of operationalization. The main point of contestability, and how we would desire to present it, is that it provides a chance for realignment.

3.4.1. CONTESTATION AND FRAMING RELEVANT MATTERS

The definition of relevancy given at the end of section 3 allows us to understand why contestability should come into play when dealing with relevancy. This does require that contestation actually leads to adaptation. It is our suggestion that such mechanisms for adaption need to be front and centre after the implementation and that it is widely available, meaningful, and effective. This entails that we also look at alignment after first implementation. In this section, I propose an initial solution.

A key problem with proposing a solution is that it requires us to design and designate contextual features and goals, which may simply not align with the uncertain nature of the world. So, if the solution were to merely describe a theoretical framework that classifies what relevancy is, then it may cause the problem we wished to avoid in the first place. Designers may overshoot or undershoot in terms of our understanding of what is morally relevant in a given situation[4]. In fact, designers may limit themselves in terms of what is con-

[4] A nice example of these problems can be found in relevancy and communication [133], whereby

textually available or what is acceptable for the telos of such machinery. At the other end of the spectrum, one may want to resort to requirements and guidelines, however those need to be followed effectively and truly, otherwise one runs into the problem of ethics washing. Furthermore, guidelines are often far too descriptive - rather than prescriptive - and can entail numerous things [149]. As Whittlestone et al [150] also mention, certain AI perspectives are simply too broad and high-level to fill in the particulars. It seems both strict and lenient solutions cause another kind of version of the frame problem. In our opinion, neither a purely theoretical framework may suffice nor a mere set of guidelines. Rather, we need a practice - a taught method to become proficient with - which entices designers to think about their method of implementation in a specific way and guide the process to alignment itself.

What the frame problem already proffers to designers is the need for iteration[5]. The example given by Dennett entails that researchers need to go back to the drawing board and see that other problems pop up after implementation. Essentially every stakeholder that comes into contact with the system or is impacted by it needs to be able to contest the design of the AI such that it can morph into something more desirable [146].

The difficulty of pointing out flaws in AI systems is that these systems are overall effective. And contesting them may harm the effectivity of the model. Yet, these models are trained to work on general cases, meaning they often align well with the general cases one deals with. However, this also means that the edge-cases designers wish to avoid may not arise at first inspection. In simple terms, reliability of a system is no necessary guarantee of safety [152]. There are bountiful examples to show that AI systems make unexpected turns and that designers lack the oversight both in its use and in the decisions such systems take [125].

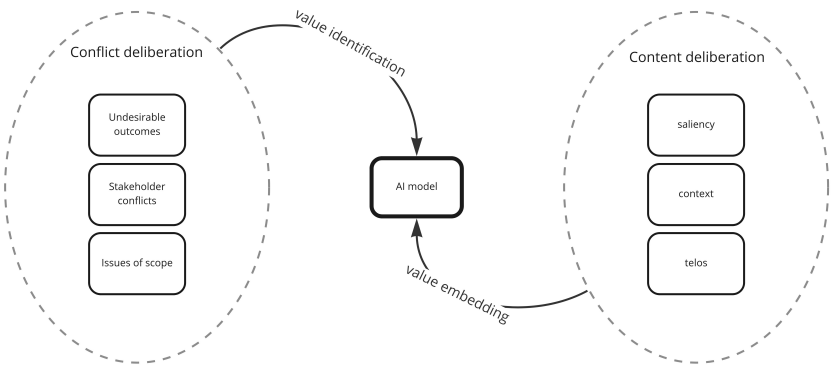
The frame problem can show designers that any process of alignment is also a process of realignment. A simple example would be a navigation system which sends users through California wildfires[153]. There is a meaningful change to the context which we need to take into account, the problem is that it isn't easy to know when this happens. Nonetheless, what is open to us is conflict, or rather: contestation. A mistake by a machine happens through conflict, of what the machine puts out and what the users need, but that is merely an indicator after the fact. Value embedding and identification, we argue, should also be conflict-driven. The frame problem shows that framing matters as we cannot put our emphasis everywhere, yet that means it will most

the authors mention the difficulty of accessing the right amount of information given a situation. The problem with a purely theoretical framework of relevance is that it invites the same limitations that we wish to avoid in the first place

[5] If we take a lesson from policy design instead, we can hearken back to Lindblom [151], he argues for the slow iterative process rather than leaps and bounds. The fact remains, Lindblom argues that in choosing policy (or in our case a specific type of implementation) is not made once and for all, it is successive because the objective and context is bound to change over time.

likely include a trade-off between certain ideals.

Thus, I believe that the constitutive elements of relevancy can feed into our ideas about what conflicts can arise during implementation and how to resolve them. These do not necessarily need to happen after a mistake has been made, but can also happen during discussions with stakeholders, the explication of implementation, argumentative structures for designing it in a particular way. In short, during each step of the life-cycle of an AI system, it could be that designers encounter a conflict. In Fig 3.1 I give a quick overview of the two sides that can go into value-alignment strategies: content deliberation (e.g. how should we formulate the context? what should we optimize for?) and conflict deliberation (e.g. different stakeholders have differing opinions). These are two sides of the same coin, as both embedding and identification of the important values rely heavily on formulation of the problem and the coinciding data collection to correctly formulate the root cause of the problem[6].



Figuur 3.1: Value identification and value embedding

Both during value identification and embedding, we can see that the constitutive elements of relevancy can be applied to think not merely in terms of accuracy, but rather in terms of outcomes and formulation. When we encounter obvious problems during this process of deliberation, we can start to understand where in our implementation certain measures must be taken. Depending on the problem at hand, a system designer can adapt one of these constitutive elements of relevancy to counteract or harmonize between conflicts. Only afterwards, when we have formulated the currently correct goal, with the correct data, and saliency, then accuracy comes into the picture as a

[6]The context of these problems are taken to be somewhat societal. These systems operate in some context that can influence or affect other agents (humans). Especially when these question become political e.g. when an algorithm which determines something about the size of the loans you can get, we inevitably arrive at the point where we must admit wickedness[120] [59].

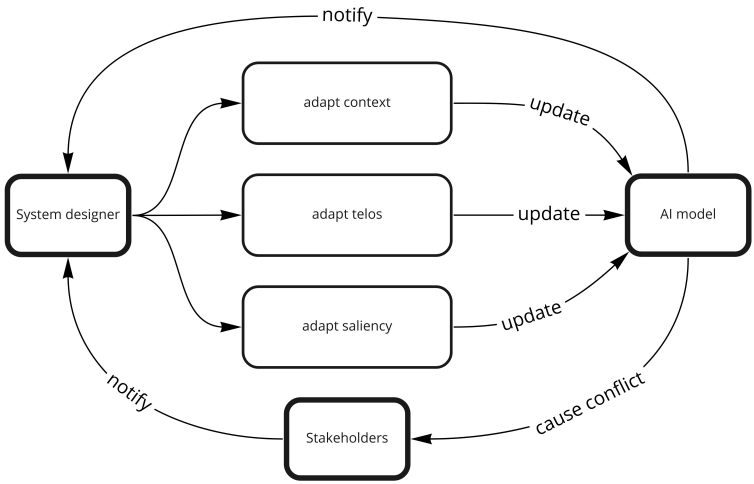
measure of knowing whether the model is effectively trained.

These types of conflict and the tools to deliberate about them don't map specifically to any element of relevancy. They are rather possible locations where the problem resides. For example, when we encounter a problem, such as sexism in automated hiring processes[154] and racially biased data for risk assessment[155], it is unclear whether we optimized for the wrong thing (telos), whether our data was limited (context), or whether it simply had the wrong means to induce patterns (saliency). In fact, all three could be the case. For example, in the case of automated hiring, one may want to weigh the maintenance of current working culture (hiring based on similarity) versus a kind of openness and serendipity (hiring based on diversity), skewing this the wrong way shows that the translation from value to goal optimization went wrong.

Yet misappropriation of goals is not the only issue. Sometimes outcomes cannot be reached anyway. The most obvious example of this is fairness, as that heavily depends on the way the problem is scoped and formulated to even begin to understand the topic of fairness in a specific contextual setting. To say that the system must be "fairïs to pull off some equivocation, because fairness to a Rawlsian may mean something entirely different to your average communitarian. Most likely, these beliefs are incompatible to such an extent that optimizing for "fairnessïs likely to result in unfair behaviour to some. To take the example of biased hiring, is sorting based on merit the way to achieve fairness? In *The tyranny of merit* [156], Sandel describes, merit may cause wildly unfair behaviour to arise. As it is bound to be influenced by socio-economic position and those who have the capacity to send their kids to all kinds of help, are going to end more or less on top. Yet, to leave it up to chance may also be unfair. We only need to think in terms of probabilities and that small percentage that never gets hired once we implement such systems en masse.

After we have derived some conflict in whatever the stage of implementation, we need to know how the conflict is embedded in the system. This even may entail that we need to change the system completely, or to acknowledge that there is no solution for the problem that accommodate such reasons in a meaningful way. As we mentioned before, it is not necessarily clear where the problem may lie within the process of embedding, it is up to the designer at the point to potentially adjust and adapt one of the three elements that we described throughout this chapter. See Fig 3.2 for a quick overview.

The figure above describes the relation between conflict and adaption. The notification can be explained through stakeholders and the AI model itself. For example, concept drift is a way of showing that a user has drifted from its original interests, therefore the model needs to be updated. However, we also propose the fact that stakeholders can do this actively, as the AI models' notification of such conflicts may not be foolproof (and open again to the frame problem). After notification of conflict, especially given the specific



Figuur 3.2: Playing the updating game

context, we do not want to look at this in an automated fashion because this may insert the original problem of the frame problem. Rather, the designer needs to play an active role in determining how to adapt the system.

Firstly, designers can adapt the context. This should be obvious, when an application is applying the wrong data, dataset, or set of propositions this can easily lead to skewed outcomes. Biased data is an often discussed topic in value alignment [157, 158]. Secondly, designers can adapt the telos, or what they optimize for. This stands in obvious connection to context. Yet, it starts with understanding that our outcomes are less descriptive and more moral than first meets the eye. Designers can view this multiple ways. Designers may gear an application towards something but leave out meaningful dimensions (like the bomb on the cart in the example of the frame problem), or they may misinterpret the situation such that these dimensions are seen as unimportant. Misfeaturization and emphasis, point us to the fact that such systems should not merely be built upon what the important facts are - but rather on what is needed and desirable here by the community at hand. Thirdly, designers can adjust saliency. For ML-approaches, this is the most difficult and time-consuming to disentangle. While in terms of relevancy this concept is about the noticeable, in AI practices designers need to understand this as induction or deduction of patterns. A particular kind of model can infer a particular kind of function, if that function disallows certain users to goods, or causes other harm, then changing the range of possibly deducible patterns may solve the problem at hand. It is easy to think about in terms of overfitting and bias.

When designers introduce more bias, any inference pattern is less likely to adapt to an outlier, meaning that the pattern induced is more interested in the general whole. Yet, the necessity of fitting should be clear, we do not all act in the same manner, or need the same outcome from a system.

For symbolic approaches, we can see that the different kinds of logics that one can use and which types of deductions are acceptable is a far quicker choice to manipulate. One could also think about what kind of inconsistencies or contradictions that result from a dataset are acceptable and how they should be resolved. The entire approach of conflict acknowledgement and adaptation through the updating of context - telos - saliency is mostly similar within symbolic approaches, except for the fact that context is often painstakingly built up from countless propositions and swapping that out is likely to be such a time-dependent and consuming project that it equates to redoing the tool.

Relevancy through contestation invites designers to think about the means by which certain outcomes are achieved and the way in which a problem can become ingrained in the system when it happens to be misaligned. This mode of thinking also shows designers a way how they can perhaps avert the problem. Just like coding etiquette, designers need to be taught in specific ways to make sure that even in larger teams with multiple designers, or working with legacy materials, they can overcome a problem of misalignment in the future. So even when the system passes hands, it is still clear why the optimization strategy is what it is, why certain contextual features are scrapped or added, and why the pattern is inferred in one way and not another. Such documentation on the possible value conflicts may allow future designers to re-align an application to the current day.

3.5. CONCLUSION

As we have seen, the openness that comes along with implementing value alignment theory can lead to misalignment during operationalization because what is relevant to a situation is far from obvious. The frame problem shows how difficult it is to get relevancy right, as it is far too easy to overshoot or undershoot in terms of deciding relevant factors. In all, we distinguish relevancy in value alignment mostly by what is thought relevant by a certain group of stakeholders, rather than say what actually is relevant. However, without means to adapt to new situations, this given relevancy is limited in a variety of ways. The context may be too limited, or the predesignation of the goal may be wrong. To effectively solve this, we argue that such systems should be built with an ingrained method to change the constitutive elements which are concerned with relevancy. We suggest that this could happen through contestability. This means that adaptability of systems needs to be in the system, together with feedback mechanisms that allow for meaningful contestation of individuals such that they can play an active part in the use of such a system.

This creates a kind of update function that may prove worthwhile in approximating a desirable outcome for man and machine.

BIBLIOGRAFIE

[15] I. Gabriel, "Artificial intelligence, values, and alignment", *Minds and machines*, jrg. 30, nr. 3, p. 411–437, 2020.

[16] F. Santoni de Sio en J. Van den Hoven, "Meaningful human control over autonomous systems: A philosophical account", *Frontiers in Robotics and AI*, jrg. 5, p. 323 836, 2018.

[22] B. Friedman, P. H. Kahn, A. Borning en A. Huldtgren, "Value sensitive design and information systems", in, Springer, 2013, p. 55–95.

[23] D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell en A. Dragan, "Inverse reward design", *arXiv preprint arXiv:1711.02827*, 2017.

[59] R. Coyne, "Wicked problems revisited", *Design studies*, jrg. 26, nr. 1, p. 5–17, 2005.

[65] L. Cavalcante Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker e.a., "Meaningful human control: Actionable properties for ai system development", *AI and Ethics*, jrg. 3, nr. 1, p. 241–255, 2023.

[100] I. Van de Poel, "Why new technologies should be conceived as social experiments", *Ethics, Policy & Environment*, jrg. 16, nr. 3, p. 352–355, 2013.

[120] R. Dobbe, T. K. Gilbert en Y. Mintz, "Hard choices in artificial intelligence", *arXiv preprint arXiv:2106.11022*, 2021.

[121] S. Umbrello en A. F. De Bellis, "A value-sensitive design approach to intelligent agents", in *Artificial intelligence safety and security*, Chapman en Hall/CRC, 2018, p. 395–409.

[122] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. Sánchez e.a., "Authors and contributors", 2019.

[123] T. Saracevic, "The notion of relevance in information science: Everybody knows what relevance is. but, what is it really?", *Synthesis lectures on information concepts, retrieval, and services*, jrg. 8, nr. 3, p. i–109, 2016.

[124] T. W. Kim en S. Mejia, "From artificial intelligence to artificial wisdom: What socrates teaches us", *Computer*, jrg. 52, nr. 10, p. 70–74, 2019.

[125] S. Russell, *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[126] ——, "Artificial intelligence and the problem of control", *Perspectives on Digital Humanism*, p. 19, 2022.

[127] B. Friedman en D. G. Hendry, *Value sensitive design: Shaping technology with moral imagination*. Mit Press, 2019.

[128] I. Van de Poel, "Design for value change", *Ethics and Information Technology*, jrg. 23, nr. 1, p. 27–31, 2021.

[129] J. F. Fisac, M. A. Gates, J. B. Hamrick, C. Liu, D. Hadfield-Menell, M. Palaniappan, D. Malik, S. S. Sastry, T. L. Griffiths en A. D. Dragan, "Pragmatic-pedagogic value alignment", in *Robotics Research: The 18th International Symposium ISRR*, Springer, 2020, p. 49–57.

[130] T. W. Kim, J. Hooker en T. Donaldson, "Taking principles seriously: A hybrid approach to value alignment in artificial intelligence", *Journal of Artificial Intelligence Research*, jrg. 70, p. 871–890, 2021.

[131] M. Peschl, A. Zgonnikov, F. A. Oliehoek en L. C. Siebert, "Moral: Aligning ai with human norms through multi-objective reinforced active learning", *arXiv preprint arXiv:2201.00012*, 2021.

[132] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman en D. Mané, "Concrete problems in ai safety", *arXiv preprint arXiv:1606.06565*, 2016.

[133] D. Sperber en D. Wilson, *Relevance: Communication and cognition*. Citeseer, 1986, deel 142.

[134] J. M. Dunn en G. Restall, "Relevance logic", in *Handbook of philosophical logic*, Springer, 2002, p. 1–128.

[135] A. R. Anderson, N. D. Belnap Jr en J. M. Dunn, *Entailment, Vol. II: The logic of relevance and necessity*. Princeton University Press, 2017, deel 5009.

[136] S. Mizzaro, "How many relevances in information retrieval?", *Interacting with computers*, jrg. 10, nr. 3, p. 303–320, 1998.

[137] A. M. Turing, "Computer machinery and intelligence", *Minds and Machines*, 1964.

[138] L. Wittgenstein, *Philosophical investigations*. John Wiley & Sons, 2010.

[139] S. G. Shanker, "Wittgenstein versus turing on the nature of church's thesis.", *Notre Dame Journal of Formal Logic*, jrg. 28, nr. 4, p. 615–649, 1987.

[140] J. McCarthy en P. J. Hayes, "Some philosophical problems from the standpoint of artificial intelligence", in *Readings in artificial intelligence*, Elsevier, 1981, p. 431–450.

[141] D. C. Dennett, "Cognitive wheels: The frame problem of ai", *Minds, machines and evolution*, p. 129–151, 1984.

[142] J. A. Fodor, "Modules, frames, fridgeons, sleeping dogs, and the music of the spheres", 1987.

[143] S. J. Chow, "What's the problem with the frame problem?", *Review of Philosophy and Psychology*, jrg. 4, nr. 2, p. 309–331, 2013.

[144] A. Hern, "Flickr faces complaints over 'offensive'auto-tagging for photos", *The Guardian*, jrg. 20, p. 2015, 2015.

[145] M. S. Pera, J. A. Fails, M. Gelsomini en F. Garzotto, "Building community: Report on kidrec workshop on children and recommender systems at recsys 2017", in *ACM SIGIR Forum*, ACM New York, NY, USA, deel 52, 2018, p. 153–161.

[146] K. Alfrink, I. Keller, G. Kortuem en N. Doorn, "Contestable ai by design: Towards a framework", *Minds and Machines*, p. 1–27, 2022.

[147] M. Almada, "Human intervention in automated decision-making", 2019.

[148] C. Sarra, "Put dialectics into the machine: Protection against automatic-decision-making through a deeper understanding of contestability by design", *Global Jurist*, jrg. 20, nr. 3, 2020.

[149] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines", *Minds and Machines*, jrg. 30, nr. 1, p. 99–120, 2020.

[150] J. Whittlestone, R. Nyrup, A. Alexandrova en S. Cave, "The role and limits of principles in ai ethics: Towards a focus on tensions", in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, p. 195–200.

[151] C. Lindblom, "The science of "muddling through"", in *Classic readings in urban planning*, Routledge, 2018, p. 31–40.

[152] R. Dobbe, "System safety and artificial intelligence", in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, p. 1584–1584.

[153] C. Olsson, "Incident number 22", *AI Incident Database*, S. McGregor, red., 2017. adres: `https://incidentdatabase.ai/cite/22`.

[154] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women", in *Ethics of Data and Analytics*, Auerbach Publications, 2018, p. 296–299.

[155] R. Richardson, J. M. Schultz en K. Crawford, "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice", *NYUL Rev. Online*, jrg. 94, p. 15, 2019.

[156] M. J. Sandel, *The tyranny of merit: What's become of the common good?* Penguin UK, 2020.

[157] M. T. Ribeiro, S. Singh en C. Guestrin, ""Why should i trust you?-explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, p. 1135–1144.

[158] B. Pratyusha, "World view", *Nature*, jrg. 583, p. 169, 2020.

4

Is Meaningful Human Control over personalised AI Assistants Possible?

You must either make a tool of the creature or a man of him. You cannot make both. Men were not intended to work with the accuracy of tools, to be precise and perfect in all their actions. If you will have that precision out of them, and make their fingers measure degrees like cogwheels, and their arms strike curves like compasses, you must unhumanize them.

John Ruskin, *The Stones of Venice*, Volume II, Chapter 6, Section 12

Recently, several large tech companies have pushed the notion of AI assistants into the public debate. These envisioned agents are intended to far outshine current systems, as they are intended to be able to manage our affairs as if they are personal assistants. In turn, this ought to give users a leg up, as one prominent tech exec has put it. Yet, it remains to be seen how these Personal AI Assistants (PAIAs) are implemented, and critical reflection on how and whether they can be implemented in a responsible way is needed. Currently, such agents are undertheorized and this may cause us to misunderstand their value and capacity. In this chapter, I explore and critique the potential for responsible implementation by considering some design requirements based on the notion of meaningful human control.

4.1. INTRODUCTION

Google, Microsoft, Meta, and OpenAI have all recently shot their developing ideas of Personal AI Assistants — sometimes simply referred to as "agents"[1] — into the public debate.[2] Supposedly, these agents will far surpass the Siris of yesteryear. Microsoft's Copilot, OpenAI Assistants API, Meta AI, Google's Gemini, and Inflection's Pi, are some of the examples that sprung up in the second half of 2023 [159]. According to an extensive report on the ethics of these "advanced AI assistants" by Google DeepMind [160] there is reason to believe that the scope of these systems will broaden, as well as their capacity for action and autonomy. While these new AI agents and personal AI assistants are still in development and the above-mentioned tech companies are vague in their descriptions of what they are aiming for, it is clear that this is intended to be the next big thing within the world of AI.

One main stated aim behind such AI assistants is that they will aid people in ordering their lives and managing certain parts of their affairs. They could also be used by organisations to increase effectiveness. As Gabriel et al. [160] explain, these agents are envisioned as being able to "plan and execute sequences of actions on the user's behalf across one or more domains and in line with the user's expectations." (cf. Gabriel et al. 2024, section 2). During an interview, Sundar Pichai, CEO of Google, said about this kind of technology that: "It may give users a leg up [161]".

The proverbial leg up here seems to relate that to what philosophers have in mind by the notion of human enhancement [162]. Yet, the discussion surrounding AI assistants can also be compared to other (AI) technology in the sense that such technology is disruptive and may alter our ways of thinking or our perception of society and morality [81]. The way in which Personal AI Assistants (PAIAs) are expected to be incorporated into our lives raises a range of additional questions. For example, these systems may be taking over tasks involving important moral and legal responsibilities [163, 164]. If these PAIAs are managing our affairs and entering into contracts for us, what will happen if things go wrong? Consequently, part of the debate ought to be about what philosophers call potential blame and praise gaps [52]. Who do we blame when such AI systems make a mistake? And who deserves credit, when such systems benefit users?

In this chapter, I am concerned with control over PAIAs. Control over such systems is highly relevant (even compared to other AI systems), given that we are handing over our personal details and personal affairs to such systems and

This chapter is currently accepted as a paper in *Philosophy & Technology*.

[1] For the sake of this paper, I am concerned with agents that resemble large language models but are likely aided by different kinds of modules for reasoning, as well as having access to different environments (e.g. financial and planning) and information. Although the term agent can mean a variety of things, the exact technical specification of what an agent is remains outside the scope of this discussion.

[2] https://www.ibm.com/think/topics/ai-agents

are supposed to let them act on our behalf in relation to (more or less important) events within our lives, both on a professional and a personal level. Meaningful human control (MHC)—a notion that has been discussed in contexts such as automated weapons systems, autonomous vehicles, and more [165] —may be regarded as a foundational requirement if we want PAIAs to be effective, trustworthy, and reliable as personal assistants. The potential for error, and possible resulting harms, is large and will be a pressing issue for the individuals using such a technology. Based on the concept of MHC, I formulate ethically informed design requirements. I apply and concretize the concept of MHC to the domain of PAIAs and show what problems we can gain insight into and what problems result from the current under-theorization of the general idea of PAIAs. These design requirements can provide a jumping off point to help shape the debate surrounding the problems that may ensue from the development and deployments surrounding PAIAs. What should we want from these technologies (and should we want such technologies at all [166])? If we desire them to be beneficial, we need to understand what trade-offs may arise from their use and whether they can actually provide the projected benefits tech companies ascribe to them.

Notably, theoretical discussions surrounding the ethics of AI assistants and AI advisors are not particularly new. Gabriel et al. (2024) suggest that such agents can help with planning, gathering information, generating ideas, finding longer-term goals, and even interacting with other humans or assistants. All these concepts have already been discussed separately in the existing (ethics of) AI literature. Similarly, before PAIAs there was already talk of AI coaches [167] and advisors [168], both of which can provide personalized help and bespoke services to people, just as skilled human assistants could. The problem of figuring out what users actually want (aligning to the right goals) has been discussed by Tielman, Jonker, and Riemsdijk [169] in their discussion surrounding decision support systems. There have also been discussions of AI assistants as potential moral advisors [170] which could aid us in decision-making and cognition surrounding difficult choices. If PAIAs will be able to enter into contracts for us, we can also take clues from automated negotiators and the related challenges, as already discussed in the literature, for instance, in terms of how to update behavioral strategies over time (as our preferences may also change over time) [171].

The potential uses of PAIAs show that we may need to deal with certain problems also established in automated negotiators, AI advisors, and coaches. Yet, we must also consider the worries that such systems may engender. This is not a completely new discussion. Several years ago, in a previous issue of this journal, Danaher [172] already proposed a general ethical framework for the ethics of AI assistants, in which he suggests that such assistants may cause cognitive degeneration or disrespect autonomy and skew our interpersonal relationships. More recently, we have seen investigations into how interaction

with AI assistants might shape our moral development [173]. Some research has also been done on the mediative standing of AI voice assistants [174] and how it may shape our values. In combining so many activities from previously separate systems (e.g. advisors, coaches, negotiators) and extending their capacities, there is a chance that new problems not seen in these different systems may arise as well.

With the proposed possibilities and extensive use in the personal domain, we need to reflect on what kind of effects these systems can have on the user and whether it is feasible or even desirable to hand over certain types of tasks to AI assistants. As Milano and Nyholm [164] already suggest, PAIAs may not be ethically or legally feasible. These PAIA systems are supposed to act on our behalf, but this may cause problems. What if the system makes a choice that affects a user who is not privy to the same information that the system has based its decision on? If an accident occurs at that point, who is responsible? And who is responsible for making sure that such a state of affairs is prevented in the first place? These systems are broadly defined, as noted above, as agents with a natural language interface, which plan and execute sequences of actions on the user's behalf but also in line with the user's expectations. But how do we know what a user's expectations are, and how personalized ought assessment to be? What types of decisions are acceptable (only low-stakes decisions or some high-stakes decisions as well?) to outsource to such a system, and which are not? As I see things, what these questions all have in common is the basic concern to enable meaningful human control over PAIAs.

Our discussion of control and PAIAs is divided into three parts. First, I briefly introduce meaningful human control (MHC). Second, I apply MHC to PAIAs and delve into three different types of issues, duties, personalisation, and idealisation, as well as the relationship between these types of issues. Issues with duties entail the problem that follows from the levels of stakes involved in the use of PAIAs. The problems surrounding personalisation has to do with how much it should be focussed on your gain versus that of others. When it comes to idealisation, I see that there is a trade-off between idealised preferences and exact preferences. In the first case it may be less comprehensible but in the latter we may not aptly capture preference drift or get stuck in suboptimal routines. After delving into the potential problems, I conclude by going over the idea of enhancing the user and investigating how this relates to the differences among users. More specifically, I analyse user enhancement from two perspectives: ubiquitous enhancement, which addresses the question of whether the improvements brought by PAIAs can be considered accessible and beneficial across different groups and actors; and directional enhancement, which focuses on whether enhancement aligns (or not) with what one truly values, taking them in their desired direction.

4.2. MEANINGFUL HUMAN CONTROL

The notion of meaningful human control originated in the autonomous weapons debate [175], with the idea that we have to ensure that humans have control of some sufficiently significant sort over advanced AI technologies intended for use in armed conflicts. The stakes in these cases are incredibly high, life-and-death decisions need to be made, and ensuring human accountability is therefore imperative. Such technology, one can argue, has to be predictable, transparent, and reliable. On the side of the human, this requires accurate information and the capacity to intervene. From those discussions spawned ideas about MHC in other domains, such as automated driving. One influential account is that of Santoni de Sio and van den Hoven [16]. They proposed looking at MHC through the notion of guidance control found in general philosophical debates about free will and moral responsibility [27]. Guidance control itself is based on the premise that people can be held responsible for actions and the outcomes they produce only if the people in question were guiding the actions in some significant sense, and this is thought to require a moderately reason-responsive mechanism and an ownership condition. What Fischer and Ravizza [27] mean by this can be summarized as follows. First, a moderately reason-responsive mechanism is a process which instantiates an action based on reasons. Thus, if different reasons apply, the action changes (the "tracking" condition). Second, the ownership condition means that the agent has taken or is able to take responsibility for the action (the "tracing" condition).

The notion of MHC over advanced technologies articulated by Santoni de Sio and van den Hoven [16] applies versions of these tracking and tracing conditions to the behavior of these technologies and their relation to the humans involved (and also changes the formulation of the conditions somewhat). As Santoni de Sio and van den Hoven [16] see things, having meaningful human control over a given system requires, first, that the technologies operate in ways that track the relevant moral reasons of the relevant people involved and, second, that it is possible to trace the behaviors of the technologies to individual(s) in the chain of events who understand the technologies and their social significance, and who can therefore reasonably be held responsible for the machine behaviours in question or their outcomes. Note that this does not mean the people in question are directly responsible at that specific juncture in time when an accident happened. A classic example of this is driving drunk. A drunk driver may not be in control over his actions, but he certainly put himself on this trajectory by drinking and getting into the car drunk. An individual can be held responsible for the trajectory they chose. Such a person may be in possession of certain agential capacities with which they could have acted differently so that they are in a general sense in control over how they act and behave in the world.

This tracking and tracing conception of MHC has recently been further

applied to other areas than autonomous weapons systems and autonomous driving [176] and can be seen as a key component of a responsible way to think about the design of values into AI systems at large [65]. Cavalcante Siebert et al. [65] suggest four defining properties based on the concepts of tracking and tracing to further delineate what the practice of MHC ought to be on the ground floor.

First, the human-AI system has an explicit moral operational design domain (MODD) and the AI agent adheres to the boundaries of this domain. Second, human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context. Third, the relevant humans and AI agents have ability and authority to control the system so that humans can act upon their responsibility. Fourth, actions of the AI agent are explicitly linked to human agents, who are aware of their responsibility.

These four properties are relatively simple to understand. The MODD criterion that is part of the first specified property implies that there is a sociotechnical element to keeping track of the relevant moral reasons. As a condition, AI systems ought to be associated with certain explicit positive and negative duties that apply within the domain in question. Even if AI systems are capable of certain actions the MODD determines that there are certain actions which ought to be undertaken and others which ought to be avoided. The mutual compatible representations suggest one part of the ownership conditions, namely that the human in question is capable of understanding under which conditions the AI system does what. From the perspective of the AI system, the condition requires that they can keep track of the limitations of the user to avoid over-reliance (which is perhaps easier said than done!). The capacity to intervene on the actions of an AI agent, as well as explicit ties to actions of the agent and human agents, allow us to tie actions of the agent to humans and makes it possible for them to be involved in some significant sense.

The following question now arises regarding the type of AI technology I am focusing on in this chapter. If we want to maintain meaningful human control over PAIAs in such a way that we can avoid some or all of the troubles described in the introduction, can we apply these four properties as design requirements? Among other things, this would require an explicit MODD, which can be comprehended by the user, such that they have the capacity to intervene if and when mistakes are made. This is, however, not as clear-cut as it seems.

4.3. APPLYING MHC TO PAIAS

So, can we—and if so, how can we—make use of the notion of meaningful human control to aid in our search for design requirements for responsible PAIAs? Before knowing whether this is the case, we need to delve deeper into what exactly a PAIA may be. Since this is still partly to be determined, what follows is speculative. Suppose that you are in some higher-up position at a large

and brilliant multinational, or that you occupy some other leadership position in which you regularly have to delegate tasks to others. You may have a human personal assistant (a PA) who helps you with your work, managing your schedule, discerning which e-mails you should read or even corresponding on your behalf, buying plane tickets, arranging visas and accommodations, carrying out background research and summarizing the findings. Perhaps they even occasionally buy gifts for some of your co-workers if you have forgotten to do that. In short, they make sure you are on time and know what you have to know at a given point and have the information ready at hand for the present situation.

Now, it ought to be clear that not all PAs are equal in terms of how well they perform. For example, they could be incompetent. In such a case, they would perhaps be sloppy, make weird inferences, mismanage your schedule, continually buy the wrong tickets, forget to arrange the correct visa (or worse, get you the wrong one), and find you subpar accommodations. They spend far beyond the expected budget. In short, managing affairs and the mistakes they make means they make your life worse off, and on top of that, increases your workload. They do exactly what they are not supposed to do. And you are likely not going to trust them with the tasks you give them. To have such a PA would require a lot of due diligence, and micromanagement just to make sure that the errors are fixed before you have to deal with the consequences. The power of a good PA has to do, in part, with taking away **your** worries and doing the correct action for **you** at the correct time. Such a PA finds ingenious ways to make your work and life easier and more comfortable, and perhaps even your decisions more informed and accurate.

The way new PAIAs are described is often in terms of making life easier. This is not surprising, of course. The companies developing these AI agents are also interested in selling and marketing their products and services, so it is only logical that they would not advertise an ineffective AI. Furthermore, who would want a personal assistant designed to carelessly do such tasks? Can we perhaps use meaningful human control as a design requirement to avoid such ineffectiveness? If we try and apply those concrete properties of MHC to PAIAs we may get something along the lines of table 4.1.

How are we supposed to introduce MHC as a design requirement, given the variance and variability of the proposed technology? As I previously mentioned, PAIAs are undertheorized. We are given a range of options to choose from regarding a level of personalization but also in terms of stakes and preferences profiles. All of these cause issues. Can we solve them with MHC?

4.3.1. ISSUES OF DUTY

The first problem that comes with trying to use MHC as a design requirement lies within the MODD. Such a MODD requires that we further detail what the explicit positive and negative duties are in a domain, but that, like the PAIA

Tabel 4.1: Applying MHC properties to PAIAs.

Original Properties	Specified for PAIA
The human-AI system has an explicit moral operational design domain (MODD) and the AI agent adheres to the boundaries of this domain.	The user needs to be informed of the positive and negative duties in line with which the PAIA should be operating.
Human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context.	The course of actions undertaken by the PAIA are comprehensible. The human in turn can inquire about said course.
The relevant humans and AI agents have ability and authority to control the system so that humans can act upon their responsibility.	The decision-making on the part of the PAIAs has to be such that the user (or someone else) can monitor/observe the decision-making and intervene if they deem it necessary to do so.
Actions of the AI agent are explicitly linked to human agents, who are aware of their responsibility.	Information about the given trajectory has to be conveyed to the human, with the consequences of that trajectory being clear as well.

itself, is underdetermined as well. Who is deciding the size of the domain? What is the extent of the context? Who gets to decide what the positive and negative duties are? Delineating a MODD for a PAIA requires that we are able to understand under which conditions and contexts such a piece of technology is implemented. Or, we could design a MODD with a specific domain and context in mind, to which the PAIA ought to adhere. Either way, it requires us to resolve contextual issues.

Consider, for example, contracts. Greediness in automated negotiation can cause harm for the individual in the long term, but it can also cause problems between individuals. If there is a large set of greedy algorithms for negotiating contracts, we may end up with a subset of users who take the brunt of the costs (e.g. the total utility of all contracts may be lower). Yet, can we enforce behaviour that leads to more optimal aggregate outcomes? In automated negotiation, contracts are notoriously difficult to design once we want a system that also keeps social welfare in mind [177, 178]. It can be done, perhaps, but what happens if an agent working towards social welfare runs into a greedy agent in a negotiation? If the stakes are low—e.g., we are merely interested in playing around with the agent for fun—then it is also easier to prefer short-term gains (ergo being greedy) for one's owner as it simply matters less. In higher stakes, this may be less warranted. Envision, an agent in the throes of

negotiating a job for you. If such agents are greedy, they may accept some job that pays well but has no long-term perspective. If we want to enforce some social welfare, then we could ground it in an explicit MODD, but this may also hamper use and effectiveness, as well as the capacity of carrying responsibility. If, for example, we consider social welfare in a generous manner, we may argue that some action was done in the name of the greater good rather than for the sake of our own benefit. We could thus wash our hands of the blame. At the same time, if they are only working for us, then they may be extremely greedy to the extent that we are all worse off, this is a classical problem in automated negotiation [171]. Greedy algorithms may work, but they may equally well create a situation in which some individuals end up victimized.

On top of that, higher stakes may make us vulnerable to losing control entirely over personal information or data. Not only do we need to ask what happens to our personal information and the company behind it that has it, such use of an advanced AI assistant is a highly desirable way to manipulate and influence individuals. If we rely on them massively, then both hackers but also the companies that can make profits from these agents' being used may want to game these agents in such a way that they can extract value from users. The problem of PAIAs in this scenario is that we give up direct control in favour of efficiency and/or effectiveness. Yet, this also means we risk losing control entirely. Di Nucci calls this the control paradox [47]. The balance to strike is one about the willingness we may have in giving up our direct control in favour of general ease. Yet, that ease may cause liability and responsibility issues and may make us vulnerable to losing control in general.

Negative and positive duties that come along with different levels of stake may cause us to consider to what extent we can rely on them. The control paradox in this case is a suggestion to what extent we may or may not want to enlarge the domain of the MODD. If the PAIA has a lot of different duties (both positive and negative) that will end up being a very different beast compared to one which is only allowed to play around in low stakes environments. The MODD will inevitably change the decision to whom to trace and what to track. In high stake environments, we will need to track different things as opposed to lower stakes. Fewer details may be relevant, and users may be more likely to take the blame due to the fact that they willingly participate in something that has few consequences in the real world. On the tracing side of things, the MODD will likely end up playing a major factor in understanding when the machine made a contextual error (i.e. went over the line of his negative duties) and thus which party to trace to.

Generally, the MODD requires knowing what the positive and negative duties are to actually implement them. This itself requires that we would know the contextual features of PAIAs. So in what environment are they nestled, whom do they interact with, what do they interact with? All of these matter to the duties in question, but for that, we need information that is (currently) not

available to us.

4.3.2. ISSUES OF IDEALISATION

A PAIA will have a range of tasks which it will need to execute well. Designing a system so that it does what you want is not as easy as it seems, as it will require knowing what is relevant to the situation [179]. Everyone makes mistakes, and the same can be said of PAs which happen to be non-human agents. In fact, the PA in question is likely going to make a variety of mistakes. Yet, the difference between a human PA and an PAIA, cannot be understated. I ame going to assume that these agents are products of companies, with whom our relationship is not the same as an interpersonal relationship with another human being. Even though we tend to anthropomorphize our relationship to technology in general, we have reasons to reign in these tendencies [180, 181]. The fact that these PAIAs are products makes them invariably different, experientially as well as legally and in terms of expectations [164, 182].

One key issue is that of representation. Should the system have an idealised representation of your wants and needs? Or is it merely a copy of your current desires (this is often called a digital twin). There is the possibility of informed preferences [15], in such a situation, the system does what I really want it to do. What this really entails is exactly the problem. It could be: what I revealed to be my preference, what I say is my preference, or what I would do if I were rational and informed. The problem is also that the machine has to translate these things into practice [125]. This translation is not a given, and neither is it certain that it actually portrays us well.

An idealised representation will likely need to capture the preferences of a user if we were to extrapolate it and include more time and knowledge. This may lead, in turn, to "better" results, however this does come at the cost of comprehensibility. It may not be needed that the user fully understands why the system is making a choice, but it needs to be interpretable. In idealised preferences, however, it would also require appropriate trust. Otherwise, it may be difficult to know when the system is going off the rails (e.g., when it is making a mistake or infers irrelevant things.). Such mistakes may undermine the feeling of responsibility towards such an agent. Even though we may have bought and paid for the service, and accepted an end-user licence agreement, that doesn't mean we will have the capacity to carry responsibility.

Idealisation may make the actions undertaken less understandable, this results in one of two things: either we would require that the human can either inquire to the extent that it isn't incomprehensible anymore, or we have to accept that there is a gap between the user's understanding of the situation and that of the system. The former is unlikely because it would require a lot of time (unless validation of preferences is very different from its calculation, but that seems unlikely) which hampers effectiveness. The other option dampens our ability to intervene. While we do not want to micromanage such

an AI assistant, the tracking condition of MHC will vary depending on whether we keep an idealised representation or not. In order to have MHC over these AI agents, we need the capacity to intervene once it is relevant. In other words, once it goes wrong. When higher stakes are involved (such as possible financial losses or risks of personal injury) we may require different capacities to intervene, we may want checks and balances beforehand. Yet, this may become much more of a hassle than we would like. Furthermore, if such systems are able to infer much more about our preferences (e.g. due to past behaviour), we also may not know whether they are doing something we reasonably desire, which could cause trouble for our intervention. For example, we could accept an outcome because we think the PAIA knows better than us, even though it is truly wrong. If such a line of thought is promising, then such a system may actively harm our choice architecture, which is the basis for what Thaler and Sunstein call nudging [183]. Theoretically, such machines can help us make better choices. If based on our beliefs, viewpoints, and arguments, those choices can help broaden our search space for optimal solutions. Yet, in effect, there is also the potential for evil by design [184], illicit use [185], or what Thaler calls "sludging" [186]. The harms of nudging vis-à-vis the possibility of autonomy have been discussed more in-depth elsewhere [185]. Yet, it is profoundly important for PAIAs to get this right. With their proposed extensive use, we really need to confront the question: Who knows what is in whose best interest? Idealisation causes issues for autonomy because it creates a lack of comprehensibility. Even with proper and sufficient interaction, there remains the possibility that we are mediated and lured towards a particular answer. If we try to counteract that by using current desires, we may diminish the effectiveness of the system.

4.3.3. ISSUES OF PERSONALISATION

Idealisation, of course, need not only be based on the individual. Idealisation can also be built upon preferences of groups rather than the individual. It can thus also be a matter of personalisation. Gabriel et al. [160], sketch a difference between personal, semi-personal, and impersonal PAIAs, by which they look at highly individualized versions of such technology and those striving towards the greater good. In comparison, in Human Compatible, Russell [125] argues that we should define the general idea of AI as artificial agents that should be working for all of us. The idea of personalized PAIAs are those that built only on your preferences, whereas the other end of the spectrum is the idea that they should work in favour of all of us. In the middle, we can think of a PAIA working for a (small) group of people.

Both tracing and tracking require that we know to which extent the technology is personalised. If it is highly personalised, we may argue that the user is more at fault, as they mould the actions more so than if it is less optimized for an individual. Faulty behaviour in unpersonalised technologies may be

more or less the fault of the designer instead, who picked an unsuitable behavioural pattern for the situation at hand. With tracking, we also need to worry about the personalisation, as this will alter the relevant details to take into account. For example, how will such a machine capture preference drifts? The mechanics of that will determine how preference profiles may change over time. In highly personalised systems, preference drift will be personalised but in impersonal systems that may be drifting in terms of a population or even be static. What is traced and tracked matters depending on personalisation.

Furthermore, personalisation may not be similar in terms of our preferences. For example, we could be in a situation where a PAIA is excluding plane tickets from your travels because a majority stakeholder (that being other people) outweighs your desire to be somewhere in little time. Such a piece of technology might be under meaningful control, but perhaps not under our own meaningful control. For MHC we merely require someone to be in control, not exactly the user, but where we end up putting control has a major influence on how we will design these systems. Which will in turn decide what we require to correctly figure out tracking and tracing. Individuals will disagree on what level of personalisation is preferable and what the greater good may be, which returns us to the point of idealisation (what would informed preferences actually be?) and the MODD (What would these duties entail?). Now, on the surface, this may seem like just another value trade-off. Is this not just another way of postulating the conflicts between the plurality that comes along with values and their alignment? We could view the situations as though we were simply debating whether to prefer one value over another [187, 188]. However, considering that these systems will likely have more application compared to the current LLMs and will do more tasks, this means that the kind of value conflicts will penetrate many more spheres of daily life than ever before. It transcends the question of value conflicts because we likely will lose the ability to detract from such PAIAs.

If our preferences about these systems differ between individuals, we would also need to address that. Our capacity to understand a representation or our capacity to intervene is certainly not equal for all. A major problem for the implementation of technologies like PAIAs, is not merely because the technology is ill-defined, but is also in no small part problematic because we differ as individuals. This is technology for somebody, and that somebody is definitely not always the same person nor are they embedded in the same context. We are simply not all made equally in terms of our preferences and our use of such systems is likely dependent on our character, temperament, cultural background, intelligence, knowledge, and environment. So what access and for whom such systems may be, is highly dependent on who we are, and that causes an additional problem in design. It is not just a value conflict, with the purported extensive use, and even adjacent use (other people using it may inevitably impact our way of life as well), instead we run the risk of throwing up

barriers for people. Before delving into that, there is one more thing we need to discuss.

4.3.4. COMPOUNDING ISSUES

I have discussed issues of duty, idealisation, and personalisation. Yet, all of these are linked. As mentioned, it is likely the case that negative and positive duties may alter based on the stakes, as well as the way that we conceive of its decision process. For example, if such a piece of technology works based on a highly informed preference profile of the user rather than their current drives, it may be required to explain to the user why a certain choice was made. Otherwise, they would not have the capacity to intervene in a meaningful fashion. Yet at the same time, this may not be possible due to the number of choices made. If, rather than informed preferences, the PAIA acts solely based on a current preference profile, then the requirement for information might be lower—as the behaviour might be more insightful—and this could create a difference in requirements for explainability. Delineating these issues makes it easier to comprehend them, but they do compound. Different preference profiles change the requirements of duties, as do stakes, as do personalisation. A high stakes environment with an idealized AI assistant might make choices and contracts which you as an individual may not want, even though they could be good for you. Idealisation and personalisation are also conjoined. A highly impersonal system may have different capacities for idealisation than one that is purely personalised. Again, the example of plane tickets would work well. Personalisation and idealisation work on different angles of the same topic, namely decision-making, but they do interact. An idealized preference profile of yourself will be different in a personalised situation as opposed to an impersonal situation. The idealisation would basically be different based on whom to take into account. But still, a digital twin that acts impersonally on your behalf is likely different from one that can take into account more time and knowledge.

Views on agency also play a contributing factor with all these issues. If we look at PAIAs through say a more agent-modelling approach (Shahom, 1997; Bradshaw, 1997), then we see that the relationship between PAIAs and agent approaches has in part to do with the proposed level of autonomy, reactivity, proactivity, and sociality. Yet, this means their analysis also has to happen more in the world (Yu, 2001) as their actions are far more dependent on worldly circumstances. The fact remains that PAIAs have to relate to worldly things, as well as being bound to epistemic limits of what can be modelled about the world. Considering that not everything in the world can be modelled, these issues described above are going to be paramount when thinking about what exactly to model and what an agent will be able to interact with in terms of domain and stake.

Nonetheless, all of these issues differ between individual users. Which

context applies? What preferences do we have? What should be the size of the domain? And furthermore, how do we take into account the fallout of use for adjacent individuals? If we join these issues together, we may need to account for the fact that not all users are equal. They may have different informational requirements, and may require more explanation of the consequences that a certain trajectory has. This, in turn, may open these users up to more detrimental manipulation or require more trust as opposed to other users.

4.4. ENHANCEMENT AND MEANINGFUL CONTROL

As we saw, there are problems with providing design requirements because of under-theorization. As a result, can we still say that this kind of technology provides a kind of enhancement? As Google's CEO mentioned, this kind of technology: may give users a leg up. This idea of a leg up reminds one of the debates surrounding Human Enhancement and how far our interference can go in adapting human lives and improving our capacities [162, 170, 189, 190]. To what extent should we improve ourselves? The obvious historical approach is one best relegated to the past, but was about improving the human gene pool. Yet, the debate on human enhancement is also about who is going to benefit [191]. It may even end up with human obsolescence [172], in which this enhancement can interfere with the skills others have taken years to learn. Copywriting comes to mind in the recent age. And given certain societal rat races, it may alternatively mean that control is lost, simply because one has to keep up with the digitally savvy Joneses. We should not underestimate societal pressures and their relation to control. One can be pressured into using technology, even if one would rather not. The question is whether MHC can combat this, or whether this is outside the scope of MHC entirely. Ideally, we create technologies to solve a problem or to improve our lives. Like freedom, technology can be a very good horse to ride, but we do need to ride it somewhere. It is this drive towards something which is currently lacking with technology like PAIAs. As Gabriel et al. [160] mention, such systems may have a profound impact on our lives. They do discuss a variety of topics to show as much. For example, they discuss value alignment, and the potential for people with higher socioeconomic status to derive more help from such assistants. They are also aware of the social impacts and the sociotechnical systems that may arise from the introduction. But the question remains: To what end?

The relation between the "end" and MHC is essential if we want to understand the problems discussed thus far. If we want to track relevant moral conditions, we do need to know what we need to track. Otherwise, we may end up aligning with the wrong reasons. If technology has to enhance us, it also has to align to our reasons and to know those reason we do require contextual features in order to model those reasons.

With all of this in mind, we need to distinguish the problem of a leg up in

two ways. Is it an enhancement for all? And is this taking us in a direction we desire? While the two are joined in some regard, it is helpful to separate these issues. It can be the case that such a piece of technology is more helpful to some than it is to others, or that it may even create a barrier. Moreover, while it may be helpful, it can push you in a direction you may not want to go. The benefit of talking about enhancement is that it presses us exactly on the question: What do we want from such technology and why?

4.4.1. UBIQUITOUS ENHANCEMENT?

Can this technology be ubiquitously good for all? Can we create technologies that enhance everyone in society? In this context, it is worth noting that Russell suggests in Human Compatible that we should define AI as technologies that solve all our problems. Yet, he realizes that we will have clashing preferences [125, 192]. o how should we understand this? Well, this starts with the obvious group that will have different values, those who are avidly opposed to the oncoming of such technologies in any way shape or form. There have been staunch defenders against versions of technology, so with their inclusion, it is already not an enhancement for all. Gabriel et al. [160] mention the values and goals of three different stakeholders: users, developers, and society. It should be clear that there may already be different incentives and conflicts between these three groups. We merely need to point back to the greediness mentioned 3.1 to see that social welfare can be opposed to certain gains for individuals. Enhancement, for all groups involved, is simply not immediately obvious.

Moving beyond these groups, we need to wonder what it means for such a technology to be good for all. Does it mean it has to improve everyone's life equally? Or merely that it improves everyone's life? If such an amendment to our personal lives is unequally distributed, it may not be seen as an enhancement. Consider, for example, a PAIA service that is offered by some tech company, which works for multiple people. What if these PAIAs do not work as well for you as they do for others? It could be that such a piece of technology is not only variable in what it can do, but also for whom it can do it and to what extent it can do these tasks well. Does MHC not imply something about the measure of control we have, or should have, over technologies? That also seems to imply something about the influence we can exert over technologies that we interact with or that make decisions for us (and about us). If formulated this way, then MHC still makes sense even in this larger societal context. In terms of the tracking condition, the PAIAs may track some users' reasons but clash with other users' reasons - and so a conflict arises about who should have MHC over the PAIA service(s) offered.

The reason why these systems may differ in effectiveness has much to do with the issues posed. Think about the differences between people. It could very well be that you have higher standards or that you request different things of technology. This would require that the MODD is particular to you, as op-

posed to what the average might be. Considering the potential personalised nature of PAIAs. We may simply run into the issues that non-personal PAIAs mean we are at a loss of control, since they act for us in a particular manner, but the same can be the case for personalised PAIAs. It is heavily dependent on what reason(s) are tracked and why. For example, if I am a highly private person compared to you, I may not be willing to give as much of my data away. Thus, if we are all forced into the ecosystem of using such technology, this may not be an equivalent upgrade for all.

There are also issues of idealisation. Consider, for example, if one has beliefs which cannot be represented by a machine (e.g. very holistic beliefs) This would mean that comparable representation does not correspond as well as it does for another individual. Does this mean that simply because one's beliefs, one's use of PAIAs is incomparable to that of your neighbours? Or you may belong to a group who is embedded in a particular context, age, or something else which hinders the PAIA from doing good work for you. It could even be a way for us to systematically deprive a group of people. An obvious example may be privacy conscious individuals, who for belief or other reasons are not willing or allowed to share their information (cf. Gabriel et al. 2024, section 13). Perhaps you don't grasp the full capacity of the machine and thus are unable to use it well. Regarding MHC, such a lack of understanding may also change our capacity to intervene, does that relate to our capacity to carry responsibility?

Other environmental factors may also impact the effectiveness. We could, for example, envision such technology interacting with other pieces of technology (IoT). If my smartwatch has an interaction with my PAIA, it may use some biodata as well to adapt one's preferences on the fly. Yet, I may not want those kinds of technologies in the first place. Another factor is whether one has a natural inclination to resist authority, a desire for more direct control, or, as we mentioned before, the desire for privacy. All of these impact both the capacity for ownership over the technology and the capacity for having the right representation. All of this in turn may affect the kind of stakes we should allow these agents to play with. What ought to be adamantly clear is that such technology will differ in effectiveness among different individuals and groups. The simple way of understanding this is in terms of information. The digital divide may grow as the complexity of technology grows too, this means that having the capacity to wield such implements and understand artefacts may be important to get to parts of society. The negative framing of this is that we are in a process of both creating effective tools for some and effective barriers for others. The embedding of MHC requires multiple trade-offs that make it hard as a design requirement, but that stems from the fact that such PAIAs are undertheorized. We can thus also view these investigations as a sort of jumping off board, to understand what problems we might run into and what situations we may want to avoid.

If we desire ubiquitous enhancement, we would thus need to proffer different kinds of personalisation tools to different people, and amend their ways on different levels. For example, if we want it to benefit everyone equally, in what way do we benefit? If it is access to the same tool, we may inadvertently help those with more social capital and time to learn to use these tools as opposed to those who don't or can't. If it should benefit everyone equally, we may need separate tools for particular individuals to improve their lives more so than that of others. Both options seem to carry with them certain benefits and certain burdens that are hard to concretize without making choices about what to prioritize. Nonetheless, if we accept that these systems can profoundly impact us, such choices may also be seen as profoundly meaningful.

We need to grapple with the fact that this enhancement could cause a widening of the digital divide. An alternative may be to impose limits on those with additional skills or access or time, or even to provide additional benefits to those with fewer capacities and worse access. At the same time, skewing it that way may be unfair and it may also mean such extensive use allows for a heightened potential for error, and a greater loss of control. Besides, it is likely that those with time and money would want the best access. Putting a limit on them would be counterintuitive to the kind of market economies we have, and would likely require heavy regulation in the first place.

If we want to avoid inequality of enhancement, we require a delineation between those who fit some preconfigured mold and those who don't. That in and of itself can pose a serious threat to meaningful control because it would likely mean that certain things will be easier to control for some than for others. A standard for MHC in such technology may thus have the opposite effect, namely a loss of control, for some individuals. We could go further and even include cultural differences between individualistic cultures and more group oriented cultures, which may impact the effectiveness of PAIAs for entire cultures as well.

4.4.2. DIRECTIONAL ENHANCEMENT

Enhancement will be very ineffective if it is bringing us to the wrong place. That being, it enhances something that we don't want at the cost of a thing we cherish. If we want meaningful control, that also means that the trajectory is actually the one we want. Consider the problem in the aid of planning. We could have a PAIA which aids us in the creation of a perfect planning on a 10-minute basis. This could be highly efficient and effective, but it may, generally speaking, not be what we want. Enhancement in this sense is directional.

The previously mentioned under-theorization matters immensely because we may not all want to apply these systems in similar situations, nor do we want others to apply them in certain capacities. This is the reason why these issues are interlinked. We interact with others and if the standard of interaction changes (say less personal due to a heightened use of assistants)

we are invariably affected even if we aren't the ones using said assistants. Yet, with regard to duty, idealisation, personalisation, we need to account for the fact that in which situations an assistant is applied and what it is allowed to do may make our lives better or worse. Some of us may not want to use it for the generation of ideas or content because we like to do it ourselves. Some of us may not enjoy outsourcing mental work because it may lead to a version of cognitive degeneration. At the same time, we might want to outsource certain tasks that we do not find enjoyable. What those tasks may be, will differ between individuals.

So what counts as enhancement? If one enjoys washing their clothes by hand, is a washing machine an enhancement? Well, it could be, if the surrounding sociotechnical system doesn't punish those who enjoy washing. If they are given ample time to do it in the way they enjoy and their life isn't altered in a meaningful fashion, it may be seen as a general addition to their life. If, however, there is societal or economical pressure to do something else, we may call this out as diminishment, at least for that individual. They may stop doing it out of pragmatic reasons because it is too time-consuming, even if they would prefer doing so it would carry with it certain costs because we have invented something new.

Now, should one individual's detriment stop the entirety of such an enterprise? Not necessarily, but we should take it into account, especially if it is not entirely clear what we are enhancing. Even if people do not enjoy a particular action, and we make it easier to do or access for many, it remains a question of what we are enhancing (and thus also for whom we are doing it). It does not seem unreasonable to say that certain unappealing activities may have benefits to individuals. In other words, the status of what actually counts as enhancement is not immediately obvious, which is why it helps to know what we are altering with technology and why we are doing it. This way we can help make better guesses who it may aid and who it might hinder.

The argument above is not meant to merely protect the status quo [189, 193]. Rather, if we devise an alternative through technology, we run the risk of creating a disability through social means. If PAIAs end up enhancing our cognitive capacities, then we are also basically creating a disadvantage for those who reject their use. And as Savulescu and Kahane point out, if we want to achieve enhancement, we do need some reasonable opinions on difficult topics like well-being and the good life.

The issues of duty, idealisation, and personalisation matter because the way we view context, how informed or revealed preferences affect us, may not be similar to all. But more importantly, the amount of personalisation may be essentially contested. It is debatable whether we will see eye to eye regarding personalisation, as this hinges on a notion of freedom weighed against a more communal notion. These differences are not only important in terms of barriers, but also what kind of society and interactions they end up shaping. As

technology is a mediating force, we need to consider whether said mediation creates a desirable society. If we simply argue that this is an enhancement period, we do not acknowledge that it may be a trajectory we do not all agree with.

4.5. CONCLUSION

It is likely that the new AI assistants of tomorrow will have broad and general capacities. Currently, however, we've seen that such personal artificial intelligent assistants (PAIAs) are under-theorized. Which leaves many important aspects of design and deployment open to interpretation. If we desire such assistants to be feasible, we will need to consider additional design requirements. What we've provided here is one such attempt. A kind of jumping off board in hopes of providing new ways to look at such technology. I consider that PAIAs will need to take into account different possibilities of both stakes and stakeholders, as well as preference profiles of individuals, and the contexts in which they may be nestled. The problem that comes from under-theorization, however, is that it is very hard to know how to build design requirements. Our application of meaningful human control still has too many open-ended questions. We may not know how to operationalize a MODD, we may not know to what extent we should include idealised preferences, nor do we know to what extent we need to personalize it. All of these choices will have major consequences on how such systems will function and how well they will function for certain individuals. Even though these systems are potentially an enhancement for the user, as long as they are aligned with the desired direction of the enhancement in terms of what one wants. We also need to know what they are meant for. We need to make sense of what is being enhanced and at what cost, and who is benefiting of those - while remaining aware that achieving ubiquitous enhancement for all groups involved might not be feasible. Otherwise, it may be likely that this kind of enhancement is a barrier to some, and it may lead us to places that we do not desire to go in the first place.

BIBLIOGRAFIE

[15] I. Gabriel, "Artificial intelligence, values, and alignment", *Minds and machines*, jrg. 30, nr. 3, p. 411–437, 2020.

[16] F. Santoni de Sio en J. Van den Hoven, "Meaningful human control over autonomous systems: A philosophical account", *Frontiers in Robotics and AI*, jrg. 5, p. 323 836, 2018.

[27] J. M. Fischer en M. Ravizza, *Responsibility and control: A theory of moral responsibility*. Cambridge university press, 1998.

[47] E. Di Nucci, *The control paradox: From AI to populism*. Rowman & Littlefield, 2020.

[52] S. Nyholm, "Responsibility gaps, value alignment, and meaningful human control over artificial intelligence", in *Risk and responsibility in context*, Routledge, 2023, p. 191–213.

[65] L. Cavalcante Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker e.a., "Meaningful human control: Actionable properties for ai system development", *AI and Ethics*, jrg. 3, nr. 1, p. 241–255, 2023.

[81] J. Danaher en H. S. Sætra, "Mechanisms of techno-moral change: A taxonomy and overview", *Ethical theory and moral practice*, jrg. 26, nr. 5, p. 763–784, 2023.

[125] S. Russell, *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[159] J. Wiesinger, P. Marlow en V. Vuskovic, "Agents", 2024.

[160] I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez e.a., "The ethics of advanced ai assistants", *arXiv preprint arXiv:2404.16244*, 2024.

[161] H. Chamberlain. (2024). "The future of ai with google ceo", Youtube, adres: `https://www.youtube.com/watch?v=h3M4bm2EveM`.

[162] N. Bostrom en R. Roache, "Ethical issues in human enhancement", *New waves in applied ethics*, p. 120–152, 2008.

[163] F. Santoni de Sio en G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them", *Philosophy & Technology*, jrg. 34, nr. 4, p. 1057–1084, 2021.

[164] S. Milano en S. Nyholm, "Advanced ai assistants that act on our behalf may not be ethically or legally feasible", *Nature Machine Intelligence*, p. 1–2, 2024.

[165] G. Mecacci, D. Amoroso, L. C. Siebert, D. Abbink, J. van den Hoven en F. S. de Sio, *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*. Cheltenham, UK: Edward Elgar Publishing, 2024, ISBN: 9781802204131. DOI: 10.4337/9781802204131. adres: https://www.elgaronline.com/view/book/9781802204131/9781802204131.xml.

[166] M. Mitchell, A. Ghosh, A. S. Luccioni en G. Pistilli, "Fully autonomous ai agents should not be developed", *arXiv preprint arXiv:2502.02649*, 2025.

[167] F. Lara en J. Deckers, "Artificial intelligence as a socratic assistant for moral enhancement", *Neuroethics*, jrg. 13, nr. 3, p. 275–287, 2020.

[168] G. Pisoni en N. Diaz-Rodriguez, "Responsible and human centric ai-based insurance advisors", *Information Processing & Management*, jrg. 60, nr. 3, p. 103 273, 2023.

[169] M. L. Tielman, C. M. Jonker en M. B. van Riemsdijk, "What should i do? deriving norms from actions, values and context.", in *MRC@ IJCAI*, 2018, p. 35–40.

[170] E. O'Neill, M. Klincewicz en M. Kemmer, "Ethical issues with artificial ethics assistants", in *The Oxford Handbook of Digital Ethics*, Oxford University Press Oxford, 2022, p. C17.

[171] T. Baarslag, M. Kaisers, E. Gerding, C. M. Jonker en J. Gratch, "When will negotiation agents be able to represent us? the challenges and opportunities for autonomous negotiators", International Joint Conferences on Artificial Intelligence, 2017.

[172] J. Danaher, "Toward an ethics of ai assistants: An initial framework", *Philosophy & Technology*, jrg. 31, nr. 4, p. 629–653, 2018.

[173] L. K. Yeung, C. S. Tam, S. S. Lau en M. M. Ko, "Living with ai personal assistant: An ethical appraisal", *AI & SOCIETY*, p. 1–16, 2023.

[174] O. Kudina, "'alexa, who am i?": Voice assistants and hermeneutic lemniscate as the technologically mediated sense-making", *Human Studies*, jrg. 44, nr. 2, p. 233–253, 2021.

[175] H. M. Roff en R. Moyes, "Meaningful human control, artificial intelligence and autonomous weapons", in *Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Au-Tonomous Weapons Systems, UN Convention on Certain Conventional Weapons*, 2016.

[176] S. Umbrello e.a., "Meaningful human control over smart home systems: A value sensitive design approach", *HUMANA. MENTE*, jrg. 13, nr. 37, p. 40–65, 2020.

[177] N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra en M. Wooldridge, "Automated negotiation: Prospects, methods and challenges", *International Journal of Group Decision and Negotiation*, jrg. 10, nr. 2, p. 199–215, 2001.

[178] V. Sanchez-Anguix, O. Tunalı, R. Aydoğan en V. Julian, "Can social agents efficiently perform in automated negotiation?", *Applied Sciences*, jrg. 11, nr. 13, p. 6022, 2021.

[179] S. K. Kuilman, L. C. Siebert, S. Buijsman en C. M. Jonker, "How to gain control and influence algorithms: Contesting ai to find relevant reasons", *AI and Ethics*, p. 1–11, 2024.

[180] J. J. Bryson, "Robots should be slaves", in *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, John Benjamins Publishing Company, 2010, p. 63–74.

[181] S. Nyholm, *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers, 2020.

[182] K. D. Evans, S. A. Robbins en J. J. Bryson, "Do we collaborate with what we design?", *Topics in Cognitive Science*, 2023.

[183] R. H. Thaler en C. R. Sunstein, *Nudge: The final edition*. Penguin, 2021.

[184] C. Nodder, *Evil by design: Interaction design to lead us into temptation*. John Wiley & Sons, 2013.

[185] A. T. Schmidt en B. Engelen, "The ethics of nudging: An overview", *Philosophy compass*, jrg. 15, nr. 4, e12658, 2020.

[186] R. H. Thaler, *Nudge, not sludge*, 2018.

[187] M. Sutrop, "Challenges of aligning artificial intelligence with human values", *Acta Baltica historiae et philosophiae scientiarum*, jrg. 8, nr. 2, p. 54–72, 2020.

[188] T. S. Petersen, "Ethical guidelines for the use of artificial intelligence and the challenges from value conflicts", *Etikk i Praksis-Nordic Journal of Applied Ethics*, nr. 1, p. 25–40, 2021.

[189] J. Savulescu, *Human enhancement*. Oxford Univ. Press, 2009.

[190] E. Juengst en D. Moseley. (2015). "Human enhancement", adres: `https://plato.stanford.edu/entries/enhancement/`.

[191] R. Sparrow, "Human enhancement for whom?", *The Ethics of Human Enhancement:: Understanding the Debate*, p. 127–142, 2016.

[192] S. M. Liao, *Ethics of Artificial Intelligence*. Oxford University Press, sep 2020, ISBN: 9780190905033. DOI: 10.1093/oso/9780190905033.001.0001. adres: https://doi.org/10.1093/oso/9780190905033.001.0001.

[193] G. Kahane en J. Savulescu, "Normal human variation: Refocussing the enhancement debate", *Bioethics*, jrg. 29, nr. 2, p. 133–143, 2015.

5

WHO DO WE TRACE TO? ON NORMATIVE REQUIREMENTS FOR MEANINGFUL HUMAN CONTROL

In spite of all similarities, every living situation has, like a newborn child, a new face, that has never been before and will never come again. It demands of you a reaction that cannot be prepared beforehand. It demands nothing of what is past. It demands presence, responsibility; it demands you.

Martin Buber, *Between Man and Man*, p.135

In the previous three chapters, we've seen issues that come along with responsible design. Part of the reason we may want something like MHC is to attribute responsibility. How do the problems from the previous chapters relate to the notion of responsibility attribution? In this chapter, I will go over the idea of responsibility and show how it is ineffective and inconclusive as to what tracing should actually end up giving us. If we want to include epistemic components, we run into the issue that they may be too burdensome to include or that they exist on top of practices and structures such that the knowledge is public. I look at Guidance Theory as a means of understanding what tracing can be and end up with the idea that tracing can best have a normative component, as it would alleviate some of the issues not only of tracing itself but also that of wickedness and path-dependence.

5.1. INTRODUCTION

In 2024, the French justice department decided to prosecute Pavel Durov, the creator of the app telegram. They prosecuted him for the purposes of facilitating criminal activities [194]. His response was: “If you are going to do this, nobody will want to innovate."He referred to the fact that innovation may be hampered if we hold people responsible for their design. We have to free the designer of the responsibility of misuse because they cannot know how the application will be used. Therefore, they cannot be held responsible. And since innovation is important, we should let them do what they desire.

The reason why the French justice department is prosecuting Pavel Durov is due to the fact that Telegram allows you to anonymously chat in huge groups. Now, we could debate about the responsibility of a company vis-à-vis a human being in such a case. But the question that I'd rather pose in this short chapter is the following: **Are people like Durov responsible for the actions of their users, or are they not?** Let's look at this from the perspective of Meaningful Human Control (MHC). To do so, we need to keep a couple of things in mind. In chapter 2, we discussed the fact that our perception may be altered by the introduction of technology. Ergo, we may lose out on alternative worldviews because of it. In chapter 3, we went over how designers were likely unaware of all the relevant details required to solve the issue correctly. They would thus invite a particular error in their design. Lastly, in chapter 4, we detailed the leg-up problem, the issue that designers may invariably end up catering more towards a particular set of users rather than all of them.

Given all the factors of the previous chapters, the question that we can ask through the lens of MHC: who should we trace to when we are dealing with pieces of (AI) technology? Given how much of an influence a designer or team of designers may have on the implementation of a system, even if they don't themselves know how exactly because of The Collingridge Dilemma [11], is it really sensible to still hide behind the idea that designers are not responsible for the actions of the users and wielders of their given implements? Is it not exactly their choices of design, their relevancy, their ideas of a user that lead us to particular harms? We will investigate this through responsibility attribution.

In this chapter, we will delve into three versions of responsibility attribution. First we go over epistemic conditions, then we go over performative conditions, and finally, we will formulate and look into normative conditions for responsibility attribution and try to translate these to tracing conditions.

This chapter provided the basis for a paper currently under review at *Ethics and information Technology.*

5.2. RESPONSIBILITY ATTRIBUTION AND TRACING

Consider some individuals are planning something illegal through Telegram. The group gets caught by the police and is held accountable for their actions[1]. The common assumption is that this accountability is due to a previous action taken. The group of users did something illegal. While they may be guilty, there is reason to believe that people like Durov also carry part of the blame. In part because this way we can account for the societal phenomenon they helped to create, and because Durov either was in a place to prevent it from happening or to amend it after an initial incident. What we will argue for is that there are likely epistemic and even moral conditions involved for attribution to work effectively.

While we are dealing with MHC, it helps to go over the general case first, that being responsibility attribution. We will spend relatively little time on MHC, as the previous chapters have gone over that in enough depth. The point of the tracing condition is that it requires some idea of to whom we ought to attribute responsibility. That is what we will mostly concern ourselves with.

Furthermore, we will not focus on situations in which either the designer is guilty of malpractice or the user is guilty of intentional misuse of a platform. We will only come back to this at the very end. I think those situations are clear in the sense that both parties would have known that there were other actions in mind. Rather, what if we intend to do well? Yet, an accident happens. Who should we trace to and why? We want to know who is responsible, and the simplest way of talking about moral responsibility is in terms of a causal condition. So we can start there:

MRA_0: Agents are morally responsible if they, at an earlier juncture T took an action that causes problems at time T+N.

While this may seem obvious at first, we do run into problems. This is seen in both the term action and the difference in delta time. The linking of actions and consequences can be very muddled. For example, if the span of time between an action and a consequence is considerable. Like the problem of many hands [54], so too can the wheel of time make it feel as if we have washed our hands of the blame. While, in honesty, we may have been responsible in the first place. The perhaps more problematic perception of MRA_0 is what it implies about action. If we consider Durov as the sole planner and builder of Telegram (which he likely isn't), we should say something along the lines of building Telegram at time T and illegality happening at time T+N. It is logical to assume that we also need to know something about the state in which Telegram was built. If its creator wasn't even aware of the possibilities of ille-

[1] There are a plethora of examples of illegal actions we could point to. Doxxing being an obvious one, but there are worse things that also happen through channels on Telegram.

gality, what does that say for their actions? It seems that there is a question of whether someone was either in a position to be aware or actually aware at the time. If someone didn't know what they were building but did so, say in small parts with no understanding of the whole, or were forced to against their will, that would change the situation [195].

What we obviously arrive at are some epistemic conditions that seem to be at play here as well. These epistemic conditions exist for the user, but also for those who desire to hold said user accountable. The simplest would be to argue that if *S knew that doing p causes a*, and *S did p, S would be responsible for a.* However, we also need to add that we as observers *O* expect *S to know that p causes a.*

For if we rely simply on the knowledge condition of the individual, we must ask, how high must the bar be for epistemic conditions? Since we cannot ask someone to know all the possible outcomes of a given action. We must instead ask about the creator. We need to answer: How can it be that they should have known? What knowledge would have been sufficient and why? It seems reasonable to assume that a designer can take into account that their design may be misused, while they may not have been instructed about the consequences. There seems to be a practice-generated entitlement, which can lead us to say that, in an altogether different situation, you ought to have known. Even if they had forgotten, such a foggy memory still would be culpable [196, 197].

Can we find a design to demonstrate such epistemic requirements? If we look at the origins of tracking and tracing, which in part come from guidance control [27], we see something that may help us. Moral responsibility for Fischer and Ravizza is reason-responsive. Fischer and Ravizza call a mechanism moderately reason-responsive if the agent has said mechanism and is capable of recognizing reasons and reacting to them (receptivity and reactivity, respectively) [198]. I am going to frame it in relatively simple terms. Moral responsibility attribution, within the realm of Fischer and Ravizza's theory, can broadly seem to mean:

MRA_1: Agents are morally responsible if they are able to suitably respond to moral reasons as well as reason about them [199].

MRA_1 as opposed to MRA_0 has the benefit of considering that at the very least, the users must be able to respond to moral reasons and reason about them. However, this begs the question of what it means if they do not suitably respond to moral reasons. Are they at fault for not suitably responding while they ought to? Is it the agent's fault for not suitably responding to moral reasons at time T which led to a problem at time T+N? Or did they not have access to the necessary faculties or information to suitably respond to the situation at hand? Merely being able to suitably respond is a welcome addition but does not allow us to comprehend what is going on when things go awry.

So when have we put individuals in the right conditions such that we can say they ought to have known? We can take a note from K.C. Clifford's ethics of beliefs here [200–202]. The premise is simply that it is morally negligent to believe without sufficient evidence. So, when ought one to have known? When one had sufficient evidence.

$MRA_{1.2}$: Agents are morally responsible if they are able to suitably respond to moral reasons as well as reason about them. And have sufficient evidence to believe particular reasons to be present.

Well, when do we have sufficient evidence? One issue that is often discussed regarding guidance control is the ownership condition [203]. If you are the one *guiding* the action, then you can be held responsible. To guide such an action requires an understanding that such an action (Action A) is issued from one's own moderately reason-responsive mechanisms. Another part is that one can have a responsibility attitude towards the action *a*, thereby making the action one's own. The problem with this view is that the process that leads to action can be completely disrupted and manipulated, such that we cannot talk of responsibility or control [204–206]. Stump and Ovenden, for example, argue that moderately reason-responsiveness can be intercepted by a plethora of things. Stump introduces an alien mastermind, and Ovenden conjures up a kind of chip. Both make individuals act in a certain manner. In both cases, there is clear manipulation of the ability to act suitably.

For the sake of simplicity, let's review Ovenden's case. Ovenden suggests a thought experiment in which weak-willed individuals can be helped to act on their judgments at the right time with a new kind of chip. In essence, the chip implanted in the brain can give the user the strength of will to act in accordance with a reason to do X. When the reason is recognized, the implant in the brain causes the formation of intention to actually do X. The mechanism to do a certain action here is obvious, but is it moderately reason-responsive? What Ovenden does is change our response to moral reasons.

Guidance control can be undermined by altering our response to moral reasons. It can also be disrupted by altering the requirements for sufficient evidence. Ovenden's brain chip could be altered to not form an intention to do X but to provide an alteration of when we consider the situation to contain sufficient evidence for a course of action. Thereby leading us to believe particular reasons to be present and therefore to suitably respond to them. The intention to actually do X is thus formed of one's own "volition"but the ownership is clearly still disrupted.

If suitable evidence can be problematically modified, then we run the risk of inappropriately acting upon the situation because our beliefs are altered. It can be done in a very obvious manner, as the brain chip could perhaps do, but it can be altered in a less pervasive manner too. For example, if we are not

provided sufficient evidence but are led to believe that we are.

Consider the situation in which a user has put its faith in an information retrieval system. The user assumes such an information retrieval system returns either a sufficient or the best possible document. They put in prompts and get certain documents from the system, which they then deem relevant or irrelevant. At first glance, this seems invariably different from the brain chip. We aren't being forced into said action, nor are our intentions or beliefs manipulated such that we accept particular reasons to be present.

Does the user have an oversight over all the different sets of possible relevant documents, from which they can judge to see which is actually the best? Clearly, that is an intractable feat for the algorithm and the human itself. Yet, the user does act based mostly on the processes of information that they had nothing to do with. What I mean to say is that users are altered by the simple use of the system itself [13, 111].

An information retrieval system can be the lens through which we define certain documents as relevant. The problem this causes for epistemic conditions is that suitably responding and having sufficient evidence are dependent on the system in a way similar to the brain chip. What if, for example, we have an information retrieval system that is heavily biased? If an individual simply does not conceive of an alternative and is lulled into the idea that there are no alternatives, is that not in much the same way an implant, which manipulates the set of possible actions? Can we truly be called owners of such an action? We may be suitably responding to moral reasons, but our beliefs about what reasons are present have been shaped.

The point of both the brain chip example and the information retrieval system is that the notion of sufficient evidence, as well as our beliefs about said evidence, is not trivial. Thus, it seems we must provide certain requirements for sufficiency of evidence as well as our ability to form beliefs about them. Luckily, there does seem to be a difference between the brain chip and the information retrieval system. The brain chip has a clear indicator of actions and beliefs altered. With an information retrieval system, the user may still have the capacity to look for defeaters. Rather than being forced, we are instead ignorant. So, when ought such ignorance to be considered culpable [207]?

Are we to blame for failing to believe *p causes a*? In his paper on Moral Responsibility and Ignorance [207], Zimmerman, gives us insight into ignorance. Let's say I knew Durov. I helped him build Telegram, only later to find out that through that action I directly supported the actions of illegality (by upping the size of anonymous channels, for example). Yet, I did not know. Should I have known? Well, we are never in control of being ignorant. Thus, ignorance itself is not an issue. It is rather that we at an earlier point should have known something to prevent that ignorance, and we failed to live up to that fact. Yet, this sounds very much again like knowing the alternatives, albeit at a different point in time.

It is also not the case that all culpable ignorance is easily traceable [208, 209]. For example, we might have been ignorant because we failed to talk to other people about the channel size. Yet, we could also trust someone who is entirely wrong at such a moment. Was it on us because we trusted that person when we shouldn't have? Or was it on them? This again is less about actual epistemic conditions and more about what one *ought to have known*. It could even be that it is about our attitude towards the situation [210, 211]. Watson gives the account of attributability [210], that we can be responsible if we acted in our right self. Thus, we believed the person when they told us we had to drag people from a scene because that was a correct expression of who we are.

With technological devices, we also need to consider something else: an information retrieval system, much like scientific instruments, as a designed object. It is designed with a promise of reliability [212]. Thus it may be false to believe that we were in a position we should have known if they turn out to be wrong. Goldberg [212] speaks of an epistemically engineered environment when we are dealing with (scientific) instruments. The information retrieval system is a good example, as it is designed to decrease the cognitive burden of acquiring knowledge. Yet, this means that we need to wonder when we ought to accept the outcome as truth[2]. For responsibility attribution, this is obviously the combination between the knowledge condition and being in a position to know. As we need both to know when to trust the outcome, others need to be able to know that you ought to have trusted the outcome.

$MRA_{1.3}$: Agents are morally responsible if they are able to suitably respond to moral reasons as well as reason about them and have sufficient evidence to believe particular reasons are present, as well as a reliable process to account for said evidence that takes into account their right self.

$MRA_{1.3}$ can solve the issue of the brain chip and the biased information retrieval system. The process by which we account for said evidence in both cases is insufficient to accept them outright. In the case of the brain chip, we lack any possible alternative ownership, and thus we are out of luck. We are not our right selves. In the case of the information retrieval system, we could have been looking for defeaters. If we did not, this placed the responsibility with us because we ought to have done so but didn't because it was an expression of our right self, and the process by which we arrived at the evidence (accepting the truth of said system) was reliable because we had, for example, enough examples in the past to trust it.

This right self is important when dealing with epistemically engineered environments. As design and inculcation may create practice-generated en-

[2] Goldberg calls this the problem of rational acceptance, and it is tied to trusting one's own faculties as well [213, 214].

titlements, we need to wonder whether the invitation to such entitlements doesn't mean design should carry part of the blame. This also has to do with technological affordances [55]. The design of these artefacts invites us to act in a certain manner, and the question is whether we can know that to its fullest extent. If we push this further with something like mediation theory, we see that the right self is a slippery slope into accepting moral responsibility unfairly. We may have been invited and altered in a way not dissimilar to the brain chip example. Ownership may be lost even though the evidence available is sufficient and accounted for in a reliable process, and we are able to respond to it suitably. It could be that the alteration of our right self leads us to take on responsibility unfairly.

If we are asking how high the bar of epistemic requirements must be, and they boil down in part to an expression of our right selves, that also invites a kind of relativism that may insufficiently account for situations in which we want to hold people accountable even though they may argue it is not an expression of their right selves. Sufficient evidence seems obvious; beliefs and processes all seem a requirement as well. Nonetheless, all of these ask more questions about under which conditions these are suitably fulfilled.

Thus, we can likely pile on more and more requirements on MRA_1. The fact remains much the same. There will be situations in which someone has responded suitably but given the wrong pretenses, and there will be situations in which someone has responded unsuitably but can defend themselves behind the fact that they did not have sufficient epistemic access.

5.3. PERFORMATIVE RESPONSIBILITY

Thus far, we have discussed the problem of knowledge and mentioned alternative possibilities a couple of times. Those alternatives were precisely what Fischer and Ravizza tried to oppose in the first place [215]. One way to look at their theories is as an argument against the Principle of Alternative Possibilities(PAP) [45]. The PAP example is an argument against causal determinism. It holds that an agent can only be responsible for an action if they could have done something else (and they can only do something else if it is not causally determined). The point for Fischer and Ravizza is that this requirement of *doing something else*, is not completely necessary to maintain control and be responsible. It merely requires that one owns the action. The main point of MRA_1 is to show that this cannot be done solely on the basis of cognitive understanding because it is simply too difficult to conceive of. Instead, the solution that Fischer and Ravizza devise is what I would call *performative responsibility*. Fundamentally, Fischer and Ravizza project a historicist view of responsibility [216] which creates a responsibility attitude. We can take note from the quote below:

> As a child grows up, he is subject to moral education (imperfect

> as it may be). The child's parents-and others-react to the child in ways designed (in part) to get the child to take certain attitudes toward himself: to view himself in certain ways. Partly as a result of moral education, the child typically acquires the view of himself as an agent, in at least a minimal sense. That is, he sees that upshots in the world depend on his choices and bodily movements. Further, the child comes to believe that he is a fair target of certain responses - the "reactive attitudesänd certain practices, such as punishment - as a result of the way in which he exercises his agency. We claim that it is in virtue of acquiring these views of himself (as a result of his moral education) that the child takes responsibility. More specifically, it is in virtue of acquiring these views that the child takes responsibility for certain kinds of mechanisms: practical reasoning, non-reflective habits, and so forth. Ordinarily, people would not characterize a child's taking responsibility in exactly this way, but this theoretical characterization gives more precise expression to the idea that the child takes responsibility for actions which spring from certain sources (and not from others).
>
> John Martin Fischer, Mark Ravizza, *Précis of Responsibility and Control: A Theory of Moral Responsibility*, page 442

What I mean by historicist, is that the child in their example learns to take responsibility because it is educated as such (i.e. the past is a certain way). It conceives of what one must be responsible for due to the interaction with one's environment and the teacher that they have. This solves in some part the problems that Fischer and Ravizza may have with the necessity of alternatives to discuss responsibility. We can be responsible for our actions because we have gained mechanisms for this and feel responsible towards these mechanisms. If we accept the idea that such mechanisms are sufficient for learning to take responsibility for certain actions, even if those sources do not always allow us to have alternatives (e.g. habits), then this leads us to the concept of MRA_2:

MRA_2: moral responsibility can only be held by agents who are capable of reasoning about moral reasons and who have a mechanism towards which they feel responsible, through things like education and inculcation.

For one, this boils down to a very practical solution to the notion of the right self: we are responsible because we are inculcated in a particular practice and thus feel responsible for said mechanism (the knowledge condition is reduced to a practice-generated entitlement). In the case of the information

retrieval system, we are taught to trust it, both by others and by using it. The outcome provides a reasonable solution, and thus we see it as reliable. It is an attitude, which on its basis may have contained certain reasons, but does not require the reasons to be present at a particular juncture. This is good because it creates exactly the conditions for guidance control.

Furthermore, it also relaxes the demand on epistemic conditions. It means we do not necessarily need to know all the possible alternatives, but rather whether the user was suitably inculcated in the practice of reasoning about the situation. If we go back to the information retrieval systems, it would likely be the notion of verifiability. Can the user verify the source of the information given, and can they reason about its correctness? If they can't, then they may not have been inculcated into the practice correctly. Yet, they ought to have known that the information may have been biased. So who is responsible there? Who is responsible for setting up the epistemically engineered environments such that we avoid the situations in which *someone should have known*?

The design and creation of an artefact and the coinciding use is worlds away from a moral education and the inflection of a child who learns to be responsible. It is not necessarily in our moral education (thus far) that we are responsible for the actions of a machine, and it begs the question: who is designing that which one is inculcated to? Who gets to decide what you *ought to have known*? Such historicist ideas of responsibility can hold in the case of moral education for a child, but it is wildly different to suggest anything of the sort for a designed technological affordance.

At first glance, MRA_2 likely puts some of the responsibility with designers. If they did not feel responsible for their own creations (creative outlet being the mechanism towards which they feel responsible), then by virtue of the definition, they wouldn't be. Which is already quite odd and likely an unwanted effect of the theory. If we do want to assume some standard by which we can say that designers have performative responsibility, then we move it outside the realm of education and inculcation again, and back into the realm of epistemic conditions. Furthermore, one could argue, designers, like users, may wrestle with the prospect of alternatives as well. They may not see all the possible alternative technologies they could design (we saw as much in chapter 3.) We could continue and argue that they are inculcated into a practice of design, but that likely leads to an infinite regress.

Generally, once we go from epistemic conditions towards more subjective conditions (feeling responsible), we also run into issues. This likely also leads to design that promotes an (incorrect) responsibility attitude in users. In general, MRA_2 is far too open and considerably morally relativistic to be conceived of as useful in application. It is not clear where we should position design requirements such that the responsibility attitude is cultivated in the right people. The responsibility in guidance control as described by MRA_2 is

foundationally subjective. We are responsible because we take responsibility. Again, one can reasonably say this may work for certain situations in which we have a habit or are inculcated in a practice in a particular way. Yet, tracing the consequences of an artefact is a different beast altogether, especially in light of potential judicial consequences when harm to other people is involved. Such feelings of responsibility are even counterproductive, as tracing appoints individuals who *are* responsible, rather than giving them the means to *take* responsibility. They are pressed into the limelight and appointed the potential scapegoat if we are dealing with the worst case. The requirement of ownership also means that it is one's *own* mechanism which is the cause for taking responsibility. Can a designed responsibility attitude that is cultivated be our own? To hazard a guess, it is likely that both manipulation and genuine cultivation are possible, but we would need to discern between those two effectively.

The claims about subjectivity are countered by Fischer and Ravizza in terms of psychopathy. If MRA_2 holds, one who has no feeling of responsibility is blameless. This is likely a given for a successful psychopath, who is able to not feel responsible for behaviour because his response to the situation was different [27]. In the case of psychopathy, we can still say that the moral reasoning skills are present and that the inculcation has happened. However, what happens is bifurcation between moral and non-moral reasons [217]. What Fischer and Ravizza appeal to is that one has to be able to appropriately respond to the moral reasons.

$MRA_{2.1}$: moral responsibility can only be held by agents who are capable of reasoning over moral reasons and who have a mechanism towards which they feel *appropriately* responsible, through things like education and inculcation.

Fischer and Ravizza consider that a psychopath is not a moral agent because they do not have the ability to act appropriately. While it may be sufficient for guidance control to say that psychopaths are not moral agents, we would like to hold someone responsible as well. If we adhere to appropriateness for our case, then we are losing the foundational subjectivity, but we introduce some kind of epistemic requirements again (e.g. when is something appropriate?). If we want to avoid that requirement, we could perhaps look at appropriateness through a normative lens.

5.4. NORMATIVE CONDITIONS FOR RESPONSIBILITY ATTRIBUTION

We have been talking mostly about responsibility attribution and guidance control rather than tracking and tracing in MHC, but the two are clearly interlinked. If we can give reasons for how moral responsibility is attributed, we

can also point our tracking and tracing conditions to follow that process.

Some underlying beliefs of MHC seem to suggest an individualistic viewpoint. The emphasis on humans, rather than groups, implies that it will be less about systemic issues and more about individual responsible actions. The rationalistic tendencies are about information and capacity rather than reasons that lie outside the scope of what can be deemed rational. While there is merit to this, clearly there are limits to it too. It should be obvious that one way of arguing responsibility is by saying the system and everyone in it is to blame. It is users and designers and policy, as well as those who place the system, who all carry a part of the blame. Designers did not account for all the users, and certain users did not make use of their own discretionary powers. There could be complacency, and on top of all that, the system itself has begun to invite a particular downstream of both new implementations and new uses. What ensues from this is a problem of many hands, in which we still have trouble attributing responsibility. Nonetheless, some middle ground is likely more true. The rationalistic tendencies forgo many other reasons that may hold ground and matter but are not as easily expressed in terms of ratio.

Is there a way that we can find normative conditions for moral responsibility attribution? We can safely say that mere causal conditions invite questions of an epistemic nature. We've seen that responsibility attitudes are too subjective and that epistemic conditions may be too cumbersome to fully implement. This led to the idea that it has to be attributable to someone based on a contextual feature, which fit with the historicist approach of Fischer and Ravizza, but that led us to the idea of appropriateness.

If we may want to go for a more normative route for appropriateness, we could turn towards societal conditions. Santoni de Sio [218], for example, describes some preconditions for MHC to work in a system. The realm we live in is not a fair one, and the reason-responsiveness may only be governed by those in a position of power. So, potentially we could look into the preconditions for MHC to arrive at the scene, which would likely turn the endeavour more into political philosophy. Santoni de Sio argues that tracing should also be free from the social identities that may affect their performance. Meaning that those who lack privilege in society ought not have their reasons disproportionately ignored. The normativity implied in the situation in which technology and MHC are applied means that we do need to deal with power imbalances as well as issues of representation. These matter for whom we can trace to in the first place, and which alternatives are open to us. Epistemic conditions may lead to intractable questions of alternatives or otherwise to questions of why said knowledge was publicly known and ought to have been known by the user as well. The latter, which is the more concrete way of looking at the situation and actually makes tracing moderately possible, does include the normative dimension and the possibility of epistemic injustice.

MRA_3: Agents are morally responsible if they are able to suitably respond to moral reasons, and those moral reasons are in effect recognized by society.

This first shot at normativity immediately lays bare a cumbersome problem because how would we get those reasons to be recognized in society? It would require a fixing thereof and has some utopian tendencies that are bound to confound a solution. This puts us in the conundrum of whose moral reasons we ought to respond to and why, and how we would go about changing that.

To amend that, we could move towards a more social connective model when it comes to tracing and responsibility. It would be an explicit take on how we should view certain moral reasons regarding technology. In Responsibility for Justice, Young grappled with replacing the notion of an individualistic model of responsibility [80]. In the book, Young's phrasing goes into notions of who has influence over the relevant social processes, who benefits from the unjust structure, and who has the ability to exact change by drawing on existing organizations. Rather than discussing the capacities as cognitive objective qualities bearing on the individual, we could instead look at them as social capacities or privileges. The main premise is to account for the structures of injustice in society, even if those structures are aided and created by individual actors who mean well. Even if their actions are normal and well-intended, their cumulative effect may produce an unjust situation. It is the structure that then is an accumulation of such individual actions enacting on their own ideals.

If we think of path-dependence, wickedness, issues of relevancy, we see that the social structure created by individuals is precisely what may cause unjust situations, even though that is likely unintended. A good version of moral responsibility attribution may alleviate some of the structures that we unintentionally created. Thus, rather than stating one action in and of itself is reason to single someone out for blame, it is instead moral responsibility attribution that revolves around the argument that this *ought* to have been prevented. What Young argues could be stated in the following way:

$MRA_{3.1}$: Agents are morally responsible for actions (even by other agents) dependent on the influence they have over relevant social processes that influence said actions, and benefits they can gain from said actions, as well as their ability to change the reasons and capacities that other agents may have by which they form the reasons to pursue such actions.

This is a sheer and obvious departure from what we have seen in some ways because we've broken up the ties between actions and consequences. It may be that some action taken by another agent can still be your responsibility even if they were unaware of what they were doing and so were you. The point

is that those in a position to change the structures have also caused certain injustices to arise in the first place.

The benefit of this is that we circumvent the problems of the knowledge condition and the issue that beleaguered performative responsibility. In effect, it is an interpretation of feeling *appropriately* responsible about a situation. Normative conditions for tracing, especially those like $MRA_{3.1}$ seem suddenly much more amenable as well as practical. Agents are not necessarily held responsible on the basis of consequences and actions, but they are also held responsible for the structures they designed in the first place.

Of course, this does require we need to discuss benefits and what those actually are. Is this merely economical? Or do social capital and the ability to exert influence over public discourse also matter? Nonetheless, the focus on responsibility as justice puts the normative conditions forward and alleviates the problems that we have encountered throughout the previous chapters. If it is challenging to change technology because of path dependency (chapter 2), if such implementation is partially opaque because of relevancy (chapter 3) and if some stand to benefit more than others (chapter 4), it seems rather obvious that MHC can better take a normative approach to tracing, since it can aid in combatting these problems.

If you stand to benefit but also run the risk of being held responsible, then you may want the ability to undo your mistake. If it is opaque in terms of relevancy, then you may want to include various stakeholders to avoid harm. And in certain cases, it may simply not be worth it to invent something in the first place because it will cause harm to a part of the population in a way that causes too many issues.

A simpler way of thinking about $MRA_{3.1}$ is to ask the following: *who had the means of exploration? Who designed such a thing and why? And who stood to profit from the design?* If exploration causes design, which becomes entrenched, then the benefits of that entrenchment ought to carry responsibility as well, considering how hard they are to undo. In general, it is those who form the practices and inculcate us into them (the companies who design the products, or, more easily said, those who stand to gain) who are likely culpable. The tracing condition, if taken to its natural conclusion, leads only to one place: the harmonious interplay between user and manufacturer, where the manufacturer always stands at the bedrock on top of which everything is built. But this requires one more thing:

$MRA_{3.2}$: Agents are morally responsible for actions (even by other agents) dependent on the influence they have over relevant social processes that influence said actions, and benefits they can gain from said actions, as well as their ability to change the reasons and capacities that other agents may have by which they form the reasons to pursue such actions. The actions undertaken by other agents for whom someone can be held responsible, must

be done without intentional misuse.

The last sentence is a requirement for normative conditions. Otherwise, we may unfairly hold designers and companies accountable for intentional misuse. But even in those case, if companies stand idly by and let it happen without counteracting it again, they become complicit and responsible due to negligence.

Now we can come back to a few of the examples that we had. Does this solve the issue of Telegram? Well, who stands to gain from the widespread use of Telegram? And who had the influence to change it? It is clear that Durov had the ability to change it and chose not to. It is also clear that owning such a platform may have financial gains. And it is clear that such a platform has an invitation for these kinds of uses (large groups and highly anonymized.)

So, if we return to Durov's claim that such personal prosecution stifles innovation, then we simply see him disavowing to take responsibility where it is due. He stands to gain from the success of Telegram. Neither the subjective responsibility attitude nor the more cognitive approach absolve Durov of his responsibility, nor do they point towards him directly, but a normative condition would allow us to point directly to the fact that he could influence the processes that take place on the app and that he has the capacity to change it. His unwillingness to do so in the case of intentional misuse merely lays that fact bare.

5.5. NORMATIVE TRACING CONDITIONS

What does this mean for MHC? We've already discussed many of the properties of MHC [16, 17], as well as the idea of tracking and tracing. When dealing with tracing, we can think of it mainly as the property:

Actions of the AI agent are explicitly linked to human agents, who are aware of their responsibility.

Siebert et al. [17] frame this property as such because it would make actions and responsibilities explicit. They suggest logging decisions, policy, use, and design such that we can infer who has particular responsibilities and why. They eventually boil this down to a knowledge condition that a user, manufacturer, or other agent should be aware of the responsibility in their actions. Even moving further, they suggest value hierarchies and structure mappings such that we can get an overview of the design choices and the stakeholder interpretations of moral reasons at stake. The major issue with this is, as we have seen throughout both this chapter and this dissertation, due to a plethora of issues that follow from unknown unknowns (e.g. the lingering effects of path-dependency, issues of relevancy, and unequal access to technology), the knowledge condition can be undermined by arguing that even though the actions were explicitly linked, many of these effects were not foreseeable. Which

is admittedly true and likely always outside the scope of what is effectively immediately traceable, unless we start providing a more normative base of tracing.

The question remains, when should we include such normative conditions? Do they always apply? If we leave room open for the chance that they sometimes don't apply, this will be a rehash of appropriateness or sufficient evidence. The suggestion is that they always apply. Tracing, in some ways, should always include a lineage tracing back to those who were able to change the relevant details or alleviate potential injustices in the situation. If it doesn't, then blaming the individual actor may not actually address the brunt of the issue that was caused by the interaction between AI and human, due to the fact that these AI systems may also invite certain behaviour. As long as we do not give that fact its due, we are likely going to maintain particular injustices in the system. And it is exactly this fact that may prevent the tracing condition from being effective and functioning properly.

So, we can effectively aid tracing conditions by also understanding that even in situations where the link between action and consequence is muddled due to lack of awareness (which is likely still going to happen sometimes), there are players who ought to have prevented that lack of awareness. And furthermore, even though there may have been many hands at the table, there were a few individuals who held the keys to change the relevant things, to influence the processes. If the ties between actions and agents are clear and obvious and the consequences are direct, and we are not unduly influenced in a mediative kind of way, we may still desire a simpler tracing condition. However, under the condition in which we are inculcated to accept the practice of certain relevancies and are mediated by the technology we use, which is perhaps even entrenched, it is hardly in any way, shape, or form fair to suggest that those who took a certain action through an agent are truly capable of both accepting alternatives and acting on them. If we truly want to respect a plurality of moral reasons in society with the use of these kinds of technologies, it must be necessary to pursue these kinds of normative conditions within the process of tracing.

5.6. CONCLUSION

In this chapter, we asked how we could adequately solve the question of whether people like Durov are responsible for the actions of their users. Built upon previous chapters, there was already an indication that this may be the case, yet neither tracing conditions nor moral responsibility attribution seemed to be adequate to fully account for this. We investigated causal, epistemic, and performative conditions for moral responsibility attribution and found them insufficient. Instead, a normative trajectory could be added to aid the finding of the relevant and responsible individual. We offer up one route for a more normative condition, through the work of Young. In this way we de-

scribe a way in which such a normative condition could perhaps be moulded, which is mostly based on the idea of who stands to gain from the technology and who has the capacity to influence it.

BIBLIOGRAFIE

[11] D. Collingridge, "The social control of technology", 1982.

[13] P.-P. Verbeek, "Toward a theory of technological mediation", *Technoscience and postphenomenology: The Manhattan papers*, jrg. 189, 2015.

[16] F. Santoni de Sio en J. Van den Hoven, "Meaningful human control over autonomous systems: A philosophical account", *Frontiers in Robotics and AI*, jrg. 5, p. 323 836, 2018.

[17] L. C. Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker e.a., "Meaningful human control over ai systems: Beyond talking the talk", *arXiv preprint arXiv:2112.01298*, 2021.

[27] J. M. Fischer en M. Ravizza, *Responsibility and control: A theory of moral responsibility*. Cambridge university press, 1998.

[45] H. Frankfurt, "Alternate possibilities and moral responsibility", in *Moral responsibility and alternative possibilities*, Routledge, 2018, p. 17–25.

[54] I. Van de Poel, "The problem of many hands", in *Moral responsibility and the problem of many hands*, Routledge, 2015, p. 50–92.

[55] W. W. Gaver, "Technology affordances", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1991, p. 79–84.

[80] I. M. Young, *Responsibility for justice*. Oxford University Press, 2011.

[111] P.-P. Verbeek, "Materializing morality: Design ethics and technological mediation", *Science, Technology, & Human Values*, jrg. 31, nr. 3, p. 361–380, 2006.

[194] K. Saris, *Telegram's durov condems arrest*, `https://www.nrc.nl/nieuws/2024/09/06/telegram-oprichter-doerov-reageert-voor-het-eerst-op-arrestatie-geen-enkele-innovator-zal-zo-nog-nieuwe-tools-bouwen-a4864797/`, Accessed: 2024-09-11.

[195] M. R. Vargas, "The trouble with tracing", 2005.

[196] H. Smith, "Culpable ignorance", *The Philosophical Review*, jrg. 92, nr. 4, p. 543–571, 1983.

[197] A. M. Smith, "Responsibility for attitudes: Activity and passivity in mental life", *Ethics*, jrg. 115, nr. 2, p. 236–271, 2005.

[198] J. M. Fischer en M. Ravizza, *Précis of responsibility and control: A theory of moral responsibility*, 2000.

[199] D. K. Nelkin en M. Vargas, "Responsibility and reasons-responsiveness", in *Freedom, Responsibility, and Value*, Routledge, 2024, p. 37–60.

[200] W. K. Clifford, *Lectures and essays*. Macmillan, 1886.

[201] S. C. Goldberg, "Should have known", *Synthese*, jrg. 194, p. 2863–2894, 2017.

[202] B. Zamulinski, "Clifford's consequentialism", *Utilitas*, jrg. 34, nr. 3, p. 289–299, 2022. DOI: 10.1017/S0953820822000139.

[203] P. Todd en N. A. Tognazzini, "A problem for guidance control", *The Philosophical Quarterly*, jrg. 58, nr. 233, p. 685–692, 2008.

[204] E. Stump, "Control and causal determinism", 2002.

[205] C. Ovenden, "Guidance control and the anti-akrasia chip", *Synthese*, jrg. 195, nr. 5, p. 2001–2019, 2018.

[206] T. R. Long, "Moderate reasons-responsiveness, moral responsibility, and manipulation", 2004.

[207] M. J. Zimmerman, "Moral responsibility and ignorance", *Ethics*, jrg. 107, nr. 3, p. 410–426, 1997.

[208] M. King, "The problem with negligence", *Social Theory and Practice*, jrg. 35, nr. 4, p. 577–595, 2009.

[209] H. M. Smith, "Non-tracing cases of culpable ignorance", *Criminal Law and Philosophy*, jrg. 5, p. 115–146, 2011.

[210] G. Watson, "Two faces of responsibility", *Philosophical topics*, jrg. 24, nr. 2, p. 227–248, 1996.

[211] N. Levy, "The good, the bad and the blameworthy", *J. Ethics & Soc. Phil.*, jrg. 1, p. 1, 2005.

[212] S. C. Goldberg, "Epistemically engineered environments", *Synthese*, jrg. 197, p. 2783–2802, 2020.

[213] R. Foley, *Intellectual trust in oneself and others*. Cambridge University Press, 2001.

[214] L. T. Zagzebski, *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford University Press, 2012.

[215] J. M. Fischer, "The frankfurt cases: The moral of the stories", *Philosophical Review*, jrg. 119, nr. 3, p. 315–336, 2010.

[216] M. E. Bratman, *Fischer and ravizza on moral responsibility and history*, 2000.

5

[217] E. Ramirez, “Receptivity, reactivity and the successful psychopath”, *Philosophical Explorations*, jrg. 18, nr. 3, p. 330–343, 2015.

[218] F. S. De Sio, *Human Freedom in the Age of AI*. Taylor & Francis, 2024.

6

DISCUSSION

Our means (passing the test) have overtaken our ends (human flourishing). And if you talk to any archer, you'll discover that to hit a target, aiming is way overrated.

Scott Newstok, *How to think like Shakespeare*, p.15

In this discussion, I will provide a lens through which one can view both the problems discussed within the dissertation as well as future research. Said lens is a particular kind of technological determinism, namely justificatory technological determinism, and it is a mainstay in common-sense thought about technology. Said technological determinism assumes certain outcomes because of technological innovation. Yet, wickedness, path-dependence, and issues of relevance all show that such an offshoot of technological determinism cannot hold ground. Even technological determinism's weaker siblings, such as pragmatism, do not survive when confronted with these concepts. Technology cannot be a singular determining factor of society if it is society that shapes said technology. Nor can the outcome of technological innovation be a given if we have problems like the Collingridge Dilemma. Through some examples, I counteract these basic instincts, and instead I propose that in doing more research, we should emphasise the subjective much, much more.

If you've read the preface, then you may have caught on to the fact that this work is but a pebble in a pond. I have discussed a few issues in the previous chapters and shown a few alternatives when it comes to the way we can view technology, but there is so much more to do. My point has been to argue the costs of technological development, and that those issues need to be solved with serious philosophy in mind and societal changes. To that end, I have taken up this research to start looking for design requirements which may improve technological development. Outside the scope of this dissertation are some more practical implementations. I could have spent more time researching the closer links between ethics and design to see if some of these design requirements, which are already quite applied, could be made even more concrete. Additional information may have been found through empirical studies. For example, seeing what the effects of feedback loops would be in AI lifecycle and seeing what a concept like mediative responsibility would do to those in positions of power. I could also have gone more abstract. There are still relations to uncover between other philosophical contexts and the design of technology, which, I believe, are primary to many of the problems we currently see with the ever-increasing onslaught of technology. Primarily, I could have delved into the relationship between formal systems and their interaction with reality. I still think the tenuous relationship between formality, language, and application is one open for investigation, which would likely lead back to the debate surrounding mathematics in the early twentieth century. However, the tension between neither fully abstract nor applied is fitting for the scene in which this dissertation was written.

Complementing the perspective taken in this dissertation, it may be worthwhile to provide a particular kind of lens to understand why these were the things up for investigation and what must still be done. It helps to understand why these chapters about relevancy and wickedness, and entrenchment all fit together. Technological Determinism (abbreviated as TD) fits nicely with the theories within this dissertation to clarify why a more pragmatical stance would not work either.

6.1. TECHNOLOGICAL DETERMINISM

The theory of TD taken at face value, is a grand theory describing technology as the great mover of societies. Like the Great Man Theory of Carlyle [219], the idea is that there are extraordinary men (or in our case technologies) which steer and shape society. Thus, both Napoleon and the internet are so extraordinary that normal rules do not apply[1]. Classically, TD contains two parts [221]: The first is the idea that innovation takes place outside of econo-

[1] While the theory has its shortcomings, it does at least provide an understanding as to how one executive of Uber could speak of their expansion in Europe in the following way: *F*cking illegal* [220] Swear words aside, the idea that one can bend the rules and laws that govern society is precisely what these theories hold in high regard.

mic, political, and social purview. The second is the idea that it drives social change. The relationship to Great Man Theory becomes more obvious by making these two explicit because it places technology outside the normal scope (being extraordinary) and it positions it as the determining factor in shaping society.

Within this dissertation, we see a clear dismantling of this idea. Relevance (chapter 3), wickedness (chapter 2 and 3), and path-dependence (chapter 2) all show that technological development is not outside our political purview, nor is it one that drives social change. While I do admit that there is a definitive relation between values and technology, the same can be said of many other circumstances we encounter in life, such as thoughts, people, and nature. To make technology the sole determinant of the fate of nations is wildly overestimating what technology brings to the table. The claim that technology is not so deterministic is certainly not new, and TD has been waning in academic debate since the second half of the twentieth century.

Nonetheless, in our modern age, we also face the ideology of TD, as Paul Edwards calls it [222]. It is the use of technological change as a justification for social change [223]. In doing so, we decouple artefacts from their political accountability. The difference herein lies that it does not outright determine technology to be autonomous or separate from society, but rather that its outcomes are a given. *Technology made me do it,* would be the mantra. It is reflected in the idea that productivity gains are the automatic result of computerization. Or perhaps more current in today's world, that AI will increase our productivity. This ideology is more akin to the Borg complex, wherein resistance to technology is considered futile.

TD is not accepted by philosophers. In fact, it is a reason to not grapple with an author's text [224]. Yet, the ideology of TD, especially as the justification of technology as a driver for social change, is still steadfast in many places in our age. Not only by technicians who herald new technology as progress, but also by politicians who argue the necessity of technology. In this dissertation, the same antidote applied against TD can be applied to the ideology of TD as well. To assume technology is a justification for a given social change is to inadequately address path-dependency and wickedness. It is to misjudge the notion of relevancy and to let some important parts of individual unquantifiable life go unaddressed. Wickedness in particular can nullify the justificatory power of technology by merely providing a different perspective which does incorporate other social costs that coincide with said given. We can see this historically by how worker conditions deteriorated with the introduction of the weaving machine, and how phossy jaw came to be because we had match factories in which we used white phosphor[2]. The use of such a substance was toxic (and known to the owners). Yet, phossy jaw continued

[2]Phossy Jaw also known as phosphorus necrosis of the jaw, is a horrible occupational disease which makes parts of the bone in the jaw die.

to exist, and thus people dealt with preventable disablement and sometimes death of women in the 19th century [225].

The point of this dissertation is in part to argue against TD but also against the ideology of TD. Even if we do, we are often left with a more nuanced stance but a problematic one still. This tends to be along the lines of: *we can improve the technology through guidelines and notions like education.* I've certainly made similar remarks throughout this work. By arguing in favour of feedback loops (chapter 3), more normative responsibility conditions (chapter 5), and incorporation of responsibility in certain areas (chapter 2), I have also maintained that we should address certain inadequacies present in current sociotechnical systems. So why would it be inadequate? What do we ought to research still?

In part, it is still a question of how this would work in actual practice. While most of the topics are not completely theoretical and abstract, neither are they fully and concretely contextualized. It remains a question whether we actually address some of these issues of entrenchment once we incorporate mediative responsibility and interoperability. The same goes for dealing with relevancy. Do people actually contest, and does this actually lead to meaningful change, or is this still open to manipulation such that people accept technology even though they should not?

6.2. TECHNOLOGICAL INJUSTICES

This dissertation has not fully addressed such a pragmatic appeal. So, to give a bird's eye account of it, what is the problem with such an idea? Are guidelines particularly ineffective? As I covered in chapter 5, the preconditions for meaningful human control require a just society such that reason responsiveness can prevail. Similarly, we need the right preconditions for society to be met before we can talk of the just expansion of technology into the public sphere. Pragmatism, as an acceptance of the current state of affairs is, in its most cynical view, nothing more than saying: *my might is right.* Yet, even here, the same examples as for TD apply here as well. The phossy jaw, from the ladies of the Victorian era, is an example of *pragmatism.* In essence, it means a preference for the greater good, even if it is gained over the backs of the few (or the many). The factory owners knew, but economic incentive prevailed. If such an argument still sounds appealing, I would hazard a guess and say that this is likely due to an oversight of said suffering or an acknowledgment of suffering but an unwillingness to give up the benefits it provides. In doing so, the injustices are bound to be maintained in one fashion or another. Furthermore, if one does accept the technology first and then considers how to do it better, one again does not adequately address the scope and problems that follow from path-dependency.

So how do we address this? Can we not solve this by technical means? That is still an open question. I, personally, would propose a different stance

towards technology. We must stop separating the technical from the social altogether; being and thinking are not distinct, nor are objects and our use of them. One route to do this is by using Latour ANT theory. He speaks, for example, about Object Agency [226], as if the object also has a driving force. I suppose that in due course this must also mean Object Politics, which is what Adorno already does in Minimia moralia [227]. In one example, Adorno speaks of our inability to close a door quietly yet firmly because doors of fridges and cars snap shut. It is the agency of an object which takes away our freedom to act in a particular manner. He ties this to the inexorable logic of capitalism, but we do not need to go there. It is, however, a reshuffling of who gets to decide what. Where Latour proffers to say that there can be a meaningful coexistence with objects and our relationship to them (humans do not always do their proper function either), Adorno instead laments the loss of freedom. I think both are right. I would assert the fact that we play an active part in deciding what the objects are around us. Both Latour and Adorno make correct assertions. Yet, only if we play an active part in letting our world be mediated by a force designed by others can we differentiate between those two. So we are collaborators to whatever the objects conspire to do as long as we remain passive. If we do not want doors snapping shut, we should be able to amend that fact, and if we cannot, only then should we agree to Adorno's assessment. In this way, we both re-establish the mind as the primary contributive source to said mediation and also give ourselves a grip on that mediating factor, such that we can be skeptical and mindful of incoming technology.

In practice, such a solution would result in a highly skeptical view of new technology, as we would rather return to protecting norms and values (whatever those may be in our current pluralistic societies) with strict adherence before unleashing the proposed progress into the world. It is a radical stance considering our current endeavours with multinationals and would require extensive legal overhauls as well as public perception and societal norms, but the promise something like this proffers is the ability to actually curtail some negative long-term side effects of artefacts. To take an example of history in terms of protecting values, corporations used to be under watch in the 19th century, as companies had the mandate to serve the common good [228], until in England, for example, the Joint Stock Company Act of 1856 was passed, which disbanded the explicit link between state and companies. From then on, the state did not need to check the social purpose of a company. Meaning that in the past, we've had far more checks and balance before we set our sights on an innovative company. Perhaps on a more manageable scale, we can speak of such a protection of values on a personal level. This can be done by avidly disavowing technology which one believes to be harmful. This requires research to consider whether it is actually harmful, which is oftentimes difficult to do, but for that, we could refer to experts. It does mean that we require experts who can look critically at what technology means to an in-

dividual, a group, and society in both the short and the long run. I suspect that these kinds of experts are in short supply, even though we have plenty of people researching these topics.

As a result, when we start changing our use of technology collectively, the technology will follow suit. This notion can dampen the effect of the Collingridge Dilemma; even though some systems may still run amok, the unwieldy nature of such artefacts is at least diminished by the sheer desire to uphold values and thus to demolish the technology if it does not adhere to them. So why is this easier said than done? Why can we not simply go along and change this use collectively?

6.3. FORGETTING ALTERNATIVES

Aside from the practical question concerning both implementation and pragmatism, there is also a more theoretical question that we really need to address and that has partially gone unattended in this dissertation. I briefly mentioned the extinction of experience in chapter 2, but that is, I believe, a far more extensive and widespread problem than I had time to delve into. One major concern I have is the fact that we may forget the alternatives present and open to us because we've simply adapted and accepted technology. I fear that we may dismiss from the mind the idea that things could be different. The theoretical question I posited as well in the preface: *What can be meaningfully erected as an artefact?* Similarly to the limits of pragmatism as I just described, there is a question surrounding what can be meaningfully designed and make it so that we plainly understand what world we are collaborating to create.

Again, problems like relevancy, as well as the leg-up problem, show that there seem to be limits to what we can promise with regard to what we can deliver. Which has some eerie similarities to the gap between rule and execution. In general, the relationship between rule and execution and how it relates to machinery is still a topic of investigation. In accepting pragmatic solutions, we are deciding to ignore this issue even if we do acknowledge it in favour of the solutions that we do desire. Yet, it also means we enforce certain patterns of society that may be unwarranted or outright discriminatory. The weighing of benefits and costs is why dealing with pragmatism also remains an open question and why it may be hard to do it effectively.

To that extent, I have made nods towards the idea that certain parts we like to quantify may in fact be unquantifiable. We can only make ourselves think we can quantify them (and then make sure people adapt to it, thus enforcing it as the truth rather than being the actual capital T-truth). This quantifiability has the capacity to discriminate against people who may not fit that mold as well as another. I spoke of this both in chapters 4 and 5 regarding the leg-up problem and the normativity in tracing, but I fear that the acceptance of technology somewhat or way before our time is a kind of entrenched discrimination that we also have mostly forgotten about.

If this is a serious endeavour that we wish to gain insight into, then I believe it starts by creating a counterweight to the viewpoint of design that I have started calling: *the production from nowhere*. It is taken by taking a cue from Thomas Nagel's view from nowhere[229], which is the detached third-person perspective, a transcendence of the *here* to a view from nowhere in particular. The production from nowhere is a suggestion that we can solve issues by assuming some third-person, detached, objective perspective. Said perspective provides a model of the world, and we can solve all our problems within said model. The model proffers ideal conditions[3], but the solution ignores a set of things that matter as well. These are in most cases longer-term consequences, as well as unquantifiable things. I hope that we can find a kind of reconciliation between the production from nowhere and our subjective desires. That we can find a kind of space for subjectivity in our design. It has unmistakably been a constant drive of mine to show that there are things we cannot design for, things we cannot know from such an objective stance. Yet through those lacunas, individuals end up being harmed. So we need to find a stance that relates to the production from nowhere but actually is somewhere, such that it does not presuppose an idealistic society that truly carries the name *Utopia*[4].

If we introduce the subjective into design, we will, by necessity, emphasize the mind, meaning, and value, and it will likely hamper the effectiveness and use of many of the tools we build. Yet, the relentless optimism and pragmatism that seem to follow from the ideology of TD truly need to be counterbalanced by the things we've lost. To be able to do that, we must never forget that there are alternatives present, that the world can be a different place. Yet technological and technocratic societies are stripping that capacity from us and making it ever more incomprehensible to see that at the same time.

So, in short, while this dissertation does address and alleviate some problems that come along with designing technology for others and does propose some philosophical solutions by introducing both limits on the impact of technologies and placing responsibility where responsibility is due, we must also admit that technology may still enforce patterns of discrimination. Once we start arguing in favour of either determinism or pragmatism, we entrench injustices while at the same time making it harder to dissent against such injustices. At worst, I fear that we may even lose out on the ability to rebel against them at all. To that end, I propose we need far more philosophical research into the more fundamental philosophical issues that come along with designing for society.

[3] Wittgenstein's frictionless ice comes to mind when dealing with the ideal conditions.
[4] Etymological speaking: Utopia is U and topos, literally meaning no place.

BIBLIOGRAFIE

[219] T. Carlyle, *On Heroes and Hero-Worship and Heroic in History*. BoD–Books on Demand, 2024.

[220] *Frequently asked questions about the uber files*, `https://www.icij.org/investigations/uber-files/frequently-asked-questions-about-the-uber-files/`, Accessed: 2025-01-13.

[221] V. Dusek, *Philosophy of technology: An introduction*, 2006.

[222] P. N. Edwards, "From 'impact'to social process: Computers in society and culture", *Handbook of science and technology studies*, p. 257–285, 1995.

[223] S. Wyatt, "Technological determinism is dead; long live technological determinism", *The handbook of science and technology studies*, jrg. 3, p. 165–180, 2008.

[224] J. D. Peters, "'you mean my whole fallacy is wrong" on technological determinism", *Representations*, nr. 140, p. 10–26, 2017.

[225] M. Myers en J. McGlothlin, "Matchmakers'"phossy jawëradicated", *American Industrial Hygiene Association Journal*, jrg. 57, p. 330–392, 1996.

[226] B. Latour, *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford, 2007.

[227] T. Adorno, *Minima moralia: Reflections from damaged life*. Verso books, 2020.

[228] L. Davoudi, C. McKenna en R. Olegario, "The historical role of the corporation in society", *Journal of the British Academy*, jrg. 6, nr. s1, 2018.

[229] T. Nagel, *The view from nowhere*. oxford university press, 1989.

7

CONCLUSION

Hundreds of people can talk for one who can think, but thousands can think, for one who can see. To see clearly is poetry, prophecy, and religion.

John Ruskin, *Modern Painters Volume III*

We return to the man who was found guilty of eavesdropping on his wife. Way back in the introduction, we said he was guilty of said act. He intended to do it. Yet, this dissertation has hopefully clarified that the man could have only taken said action because he was given a particular range of actions. The company that mounted the tablet, likely took shelved parts which happened to contain a microphone. The regulators that created policy never forbade the unintended use of shelved parts which contained a microphone. Then there are other designers who created those shelved parts such that the tablet could be designed in the first place. It is a trace upon a trace upon a trace that leads back further into history. It contains designers who designed new pieces of technology, as well as early adopters and marketers who pushed it into the hands of the many. Said designers followed up on previous designs of other designers. Each one of them carries part of the blame for the creation of the range of alternatives. Did all of them intend to let the man eavesdrop on his wife? Surely not. Did they help shape society such that it could happen? They surely did.

What this demonstrates is the sheer malleability of regulation and design surrounding artefacts. At any point, we could have made a difference. Perhaps not as individuals, but definitely as a society. If the underlying premise of mediation theory is true, namely that technology is constitutive of what it means

to be us, then uncontrolled technological innovation is like the alteration of who we are and what we are without any kind of standard. Due to perhaps profit motives or short-sightedness[1] this will likely make us worse for wear. So what can we learn from all this? Let's first return to what this dissertation is all about.

Technological development is wicked and normative and carries a cost. To solve issues, it is essential that we look at development through a philosophical and political lens.

By now we should have a better understanding of why I wanted to research the costs of technological development and our ability to perhaps curtail these costs.

The promise of technology is easy to see, but the costs are challenging to weigh, and that makes a lot of technological innovation murky in terms of a cost-benefit analysis. I argue that normativity is essential for the proper understanding of technological development. We cannot wrest free from many of the issues that I posit. While there are, for example, debates about mediation theory or technological affordances, I instead chose to focus on the way we can perceive the essential normativity stemming from wickedness that comes along with technological development.

What we saw in chapter 2 was an embeddedness that created a new standard. We discussed how path-dependency could make it harder to deviate from paths of exploration and design as new design is built on previous design. All of this being obviously wicked. In chapter 3, we discussed the problem of relevancy, in which it was clear that we weren't in complete control of the design process. It could be that we didn't know all the relevant factors to take them into account, again, obviously wicked. In chapter 4, we further delved into both of these issues through the case of PAIAs. We also showed that it may very well be that we did not account for the wealth of diversity among people, and if you read between the lines: the lack of theorization made implementation and selling technological promises easier. If we hearken back to wicked problems, we see that this means PAIAs do not adequately address the multivariate nature of wicked problems nor take into account the different perspectives that may live between humans.

All of these chapters point to different costs, but all of them also have to do with the creation of new standards, in which some people perhaps benefit and others certainly don't. The QWERTY keyboard eventually replaced almost

[1] Myths help us in understanding such short-sightedness. A well-known example in value-alignment literature is the problem of excessive literalism, which is portrayed by King Midas and his golden touch. In actuality, Pandora is the likely candidate for the larger problems we face with technology. Her jar (or the often mistranslated box) we can represent as the promise of progression. It is an artefact, after all. Her curiosity is what makes her open the jar to unleash untold horrors into the world. Yet, it is her short-sightedness and carelessness that leaves hope still stowed away in the bottom.

every other, becoming the standard. Each time a promise is given for a new piece of technology, we also see new problems arise. Penicillin effectiveness turned out to be a factor in stopping virological research. Recommendation systems gave us easy access to information, but also skewed our search results in unintended ways. PAIAs could give us easier control over our lives, but perhaps could work to our detriment just as well. There is a normativity and wickedness involved as we make countless choices about technology. We cut corners (by the sheer necessity of using models) when we look at the problem we desire to solve, and our solutions are heavily dependent on the way we formulate those problems in the first place as well as the previous solutions we had at hand, and thus we are dependent on the way we formulate our problems. While there are plenty of costs, the problem is that those costs are displaced, only visible by people who come after us, or who we seem to have little responsibility towards.

All of this invites us to the question: can we make a better standard for technology? Can it be improved? Is there something we can do to combat wickedness in technological development? We can review the chapters through this idea as well. In chapter 2, I mention limits to scale and required interoperability of technology, as well as mediative responsibility for designers. We want technology to be limited in scale, to be easily interchangeable by something similar, and for designers to carry a part of the responsibility by virtue of designing something that is embedded. The obvious normative component is the argument that the mere possibility of having different standards of living is worthwhile. On the aspect of wickedness, mediative responsibility is what Rittel and Webber also suggest, namely: The planner has no right to be wrong. In our case, the designer has no right to be wrong because it may cause all kinds of issues for various individuals. In chapter 3, I discuss the potential for feedback loops, to counteract the problem of overlooking relevant factors. In the process, we could redefine some of the technology to better incorporate the issues we encounter. Again, the argument follows that we need to account for individuals or factors that we cannot immediately see, that may entail a different standard of living. It is a solution for some of the normative aspects, but with wickedness we still run into the issue posited by chapter 2 (namely that it may push us down a particular direction). In chapter 4, I argue in favour of looking at issues of representation. We need a kind of technological equity for those who cannot be represented well by said systems. Again, it hammers home the same argument: We may be different in what we find beautiful in life, what we value in life. We may have different standards for what a good life entails. Technology has a normative component simply because it may eradicate different standards of living.

Furthermore, there is the aspect of control. How are we supposed to deal with it? It seems to be insufficient to merely measure who is in control. We lack the relevant details to know that beforehand, and we do create a system in

which people may learn to take responsibility for something that they should not assent to in the first place. Which, in the long run and because of path-dependency, may be harder and harder to deviate from. So we may end up entrenching a performative version of control that is not at all responsible.

Even if we add meaningful human control as a design requirement to technological development, as we tried in chapter 4, then we still need to account for the obvious limits of meaningful human control. On this scale, we may need a kind of epistemic humility in design which lets us treat human beings as an end unto themselves, in a Kantian sense. But perhaps more pressingly, as Buber explains so well: indivisible wholes. We are not equivalent creatures, never truly comparable to one another, as we have a viewpoint of one. Nor can one situation be compared to the next. Our viewpoint is never truly detached and objective, as computer science seems to appeal to. Yet, perhaps we have the chance to create a kind of intersubjective approach between us, in which technology can play a role. One way to conclude this is what I also provided in chapter 5. This could go through the idea of responsibility as a mechanism to counteract social injustices, and to reshape parts of society can be beneficial as a means to hold those accountable who have created certain technological problems.

Normativity is present in technological development, that is in part why we can see it in terms of benefits and costs. But as we make more technology, it is that normativity that is becoming a larger part of how we develop our society. We've known that for a longer time. Yet, that is so essential to it, and how the costs interact with said normativity has hopefully become clearer. With more technological development comes the increasingly heavier burden of responsibility, and that needs to be placed where responsibility is due.

EPILOGUE

I'd be remiss if I did not explain the extent to which this topic reaches and the research that goes with it. Many of the things I researched did not neatly fit the process of this dissertation. I have delved into more computer scientific research, and I have explicitly delved into philosophy. Outside the realm of science, I have written for laymen and given lectures to laymen. Some of this is added in the references below; more will likely follow. I also think literature has to play a part in understanding some of the more emotional depths that follow from the radical changes we are introducing in the world. It is the reason I took so many notes from literature in my introduction. There are things that science cannot convey in a similar manner.

The ways in which the topics outside the scope helped me were quite obvious to me. Researching concept acquisition led me to new ways of thinking about understanding, which gave me insight into another view on value alignment. Interacting with students just showed me how monumental the task was to communicate different ideas from different fields. Laymen required me to close the gap between academic life and the rest of the world, which is how justificatory technological determinism came into the discussion. Conversations with philosophers showed me their lacunas in their technical knowledge, and discussions with computer scientists showed me how uninformed they were on philosophical matters. In all, these things helped inform and shape the way that I communicated about the concepts within this dissertation.

This research has made me reflect on my own stance on science. In retrospect, this has been an important takeaway for me. The best kind of science is done when it is not merely an abstract game of writing responses for other academics, but neither something that is merely intuited for practical or popular scientific endeavours. The practice I hold dear also entails that we must remain able to condemn those who provide us with pay. Even though we are likely embedded in universities, we cannot be beholden to any institute. Furthermore, science is an engagement with the world that does not need to consider the zeitgeist of the time as pertinent or relevant. While there are things pressing in the here and now, we must also be able to question even the most stable and common-sense convictions if we suspect they may be wrong. To do so, we require a kind of freedom and independence, which can be threatened from within and without. If we don't guard those freedoms, we will never be able to combat the entrenchment of bad ideas.

BIBLIOGRAFIE

[230] S. Kuilman, *Waar blijft de planologie van AI? - Computable.nl — computable.nl*, https://www.computable.nl/2022/06/21/waar-blijft-de-planologie-van-ai/, [Accessed 11-03-2025].

[231] ——, *Rot op met je foutmarge - Kaf — kaf.online*, https://kaf.online/rot-op-met-je-foutmarge/, [Accessed 11-03-2025].

[232] ——, *Onze overtuigingen moeten niet wijken voor technologie. - Kaf — kaf.online*, https://kaf.online/onze-overtuigingen-moeten-niet-wijken-voor-technologie/, [Accessed 11-03-2025].

[233] ——, *Sommige algoritmes zijn een gevaar voor iedereen - Kaf — kaf.online*, https://kaf.online/sommige-algoritmes-zijn-een-gevaar-voor-iedereen/, [Accessed 11-03-2025].

[234] ——, *Kunnen we wel spreken van 'intelligente' AI? — nporadio1.nl*, https://www.nporadio1.nl/fragmenten/de-nacht-van/9c542ed3-995e-4bbe-90ac-e14ad74526ea/2024-06-20-kunnen-we-wel-spreken-van-intelligente-ai, [Accessed 11-03-2025].

[235] S. Kuilman en L. C. Siebert, *The Frame Problem, The TAILOR Handbook of Trustworthy AI — tailor.isti.cnr.it*, http://tailor.isti.cnr.it/handbookTAI/Accountability/L3.The_frame_problem.html, [Accessed 11-03-2025].

[236] S. Kuilman, *Wicked problems, The TAILOR Handbook of Trustworthy AI — tailor.isti.cnr.it*, http://tailor.isti.cnr.it/handbookTAI/Accountability/L3.Wicked_problems.html, [Accessed 11-03-2025].

[237] ——, *The Problem of Many Hands, The TAILOR Handbook of Trustworthy AI — tailor.isti.cnr.it*, http://tailor.isti.cnr.it/handbookTAI/Accountability/L3.Problem_of_many_hands.html, [Accessed 11-03-2025].

[238] S. K. Kuilman, K. Andriamahery, C. M. Jonker en L. C. Siebert, "Normative uncertainty and societal preferences: The problem with evaluative standards", *Frontiers in Neuroergonomics*, jrg. 4, p. 1 147 211, 2023.

A

Brief Explanation of Key Terms in Alphabetical Order

For readability, I have added a brief explanation of key terms used within this dissertation.

- **The Alignment Problem.** This is a topic of debate in philosophy of technology and revolves around the inability of algorithms and AI systems to capture our norms and values. The easy way to think of this is through the often used example of King Midas. King Midas asked of the God Bacchus the blessing to turn stuff into gold at his touch, which turned out to be a curse. As a consequence, Midas couldn't eat because what he touched turned into gold. Now, what if rather than a God, an algorithm gave Midas the ability to turn the stuff he touched into gold? The same problem would arise unless these systems were aligned to us properly. We do not want such AI systems to be excessively literal; we want them to do the appropriate thing, to be aligned with our values, to infer the proper thing. For those more philosophically inclined, the alignment problem is related to the problem of the Rule Following Paradox, which is a problem in philosophy of maths and philosophy of language. It is also related to **The Frame Problem**. Another name for this topic as a whole is **value alignment**.

- **The Collingridge Dilemma.** This is a very well-known dilemma in Philosophy of Technology. It is a double bind and states, on the one hand, designers are unaware of all the **relevant** factors and the impact their

design will have until it is widely implemented. Yet, once it is widely implemented, they lack the power to change the technology because it has become **entrenched**.

– **The Control paradox.** This paradox stems from putting control in one place, which may first give you more control but also puts you at risk of losing more control. For example, you can have many different passwords, but to remember all of those is impractical. You may forget those, and if many of the passwords are the same, they may be very unsafe. It is much better to use a password manager. It would allow you to have far more complicated passwords and more different ones as well. However, now that you've centralised the place of control, you also risk losing more control. Because losing one normal password is manageable, losing all your passwords at the same time because you forgot the key to your password manager is more problematic. The paradox lies in the fact that improving our control over certain situations seems to cause our loss of control to be greater too.

– **Entrenchment.** This is a process by which things become fixed. It relates quite heavily to **self-reinforcing effects**, **The Collingridge dilemma**, and **path dependency**. Entrenchment means that the costs of changing our ways (e.g. undoing technology) are quite high because it has become a fixture. One simple example is cars. The technology is a motorized vehicle. Yet, why it is hard to remove is likely obvious. Not only because we have numerous cars already, but also roads, and people who teach you to drive, the people who make you do tests, and the signs that have been made. Furthermore, we are used to living with cars and a society in which cars are very normal. To dig cars out of the trenches takes a lot of work; ergo, the cost is very high.

– **The Frame problem.** Very briefly put, The Frame Problem entails that AI systems cannot be programmed with all the required information to avoid unwanted outcomes by virtue of the fact that they use some version of a model (statistical or logical) to interact with the world. Another way of putting it: AI systems lack common sense.

– **Guidance control.** Basically, this is the idea that you can be morally responsible for an action if we identify with the reasons that lead to said actions and for the reasons to be connected to actions appropriately and to connect those in turn with the external world appropriately. What this means to say is, sometimes we do not always have an alternative course of action to take, but because we consent to the reason for taking our current course in the first place, we can be held responsible.

– **Path dependency.** The notion of path dependency has to do with what follows from **entrenchment**. One piece of entrenched technology may

lead to the next and the next, each of them becoming dependent on the previous one. A weak version of this would be to say, The past constrains the future. This may not necessarily be a problem unless one of the things in our past turns out to create some kind of injustice.

- **Personal Artificial Intelligent Assistant.** This is a particular type of AI system that is quite new and not implemented into society. The idea is that it should somewhat emulate a personal assistant. How they are defined is as follows: artificial agents who can plan and execute a sequence of actions on behalf of some user across domains and in line with the user's preferences.

- **The Problem of Many Hands.** This is a philosophical problem that stems from the distribution of responsibility. If many people play a tiny part in an act, they are likely unable to accept responsibility for the whole. Many people may dilute the responsibility such that we have no clear attribution of moral responsibility.

- **Meaningful Human Control.** The titular Meaningful Human Control, or MHC for short, stems from a debate about killer robots. Some argued that people should always be in control over such bots in a meaningful fashion. Over time, this idea was taken up by a larger community to also consider other domains, such as self-driving cars. One prominent way of looking at it is as an implemented version of **Guidance Control**. In that case, we are looking at a **tracing condition** and a **tracking condition** to give an account of how we can maintain control and make sure it is meaningful too.

- **Mediation theory.** This theory's central idea is to approach technology as a kind of mediator between humans and others and humans and the world. It means that technology can help us see the world in different ways. A simple example would be to think of using scans to see a baby. In this way, we can learn to see babies before they are born, but we can also designate them as a patient (if something is wrong). That experience can exist mostly because it is mediated by the kind of scanning technology that we have.

- **Relevancy.** This is a topic far too large to be briefly summarised in one paragraph. Yet, we are going for a very intuitive notion here. Things that are relevant matter for some task or outcome at hand and likely influence them.

- **Responsibility Gaps.** Responsibility gaps are another prominent feature in philosophy of technology. It revolves around the question of who may be responsible once a system of high enough complexity is used and makes a mistake. It is somewhat similar to the **problem of many**

hands. To be very brief, the link between action and consequences is muddled through distance. For example, a designer may not have intended for a certain kind of mistake to happen years down the line, and thus they may not feel responsible at all. It is simply not clear how one should attribute moral responsibility to someone in such a situation.

– **Self-reinforcement.** Effects that have self-reinforcing tendencies can be easily equated to habits. Once you get into a habit, it becomes easier to do, and therefore you do it more often. In relation to this dissertation, we can think of getting used to a certain piece of technology (which is also a habit). If a majority of people know how to use a QWERTY keyboard, that means it is more likely for companies to buy QWERTY keyboards because on average most people know how to use those. Therefore, more people can use those kinds of keyboards. It has obvious ties to **entrenchment** and **path dependency**. If the things that reinforce themselves also become fixed, then we are getting relatively close to one clause of **the Collingridge Dilemma**.

– **Technological Determinism.** Technological determinism turns technology into the great mover of societies. It is the belief that the internet overturned the world, rather than our collective use thereof. Simply put, the belief in technological determinism makes technology lie outside the realm of our influence, and outside political purview as well.

– **Tracing condition.** One of two conditions for an often-mentioned version of Meaningful Human Control in this dissertation. The tracing condition requires that we are able to trace actions, consequences, behaviour, capabilities of a human-AI system to one or more individuals linked to that action who were aware of it in a technical and moral sense.

– **Tracking condition.** One of two conditions for an often-mentioned version of Meaningful Human Control in this dissertation. The tracking conditions will require of a human-AI system that it is capable of inferring the relevant moral reasons and should be responsive to them under the relevant circumstances.

– **Wicked Problems.** Wicked problems are the kinds of problems whose solutions are dependent on how we formulate the problem in the first place. If we want to combat health issues in society, and we believe that those stem from living too close to roads, we will have a drastically different solution as opposed to someone who believes the problem stems from differences in socioeconomic status.

SIKS DISSERTATIONS

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

	40	Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
	41	Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
	42	Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
	43	Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
	44	Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
	45	Bas Testerink (UU), Decentralized Runtime Norm Enforcement
	46	Jan Schneider (OU), Sensor-based Learning Support
	47	Jie Yang (TUD), Crowd Knowledge Creation Acceleration
	48	Angel Suarez (OU), Collaborative inquiry-based learning
2018	01	Han van der Aa (VUA), Comparing and Aligning Process Representations
	02	Felix Mannhardt (TU/e), Multi-perspective Process Mining
	03	Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
	04	Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
	05	Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
	06	Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
	07	Jieting Luo (UU), A formal account of opportunism in multi-agent systems
	08	Rick Smetsers (RUN), Advances in Model Learning for Software Systems
	09	Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
	10	Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
	11	Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
	12	Xixi Lu (TU/e), Using behavioral context in process mining
	13	Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
	14	Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
	15	Naser Davarzani (UM), Biomarker discovery in heart failure
	16	Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The "Kïn "semantic web"stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OU), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TU/e), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
	31	Gongjin Lan (VUA), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?"How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence

16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data