Persuasive technology, allocation of control, and mobility : an ethical analysis

Persuasive Technology, Allocation of Control, and Mobility

An Ethical Analysis

Persuasive Technology, Allocation of Control, and Mobility

An Ethical Analysis

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen op dinsdag 13 maart 2018 om 16:00 uur

door

Jilles Smids

geboren te Apeldoorn

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. I.E.J. Heynderickx
1e promotor:	prof.dr.ir. A.W.M. Meijers
copromotoren:	dr. A. Spahn
	dr. P.J. Nickel
leden:	prof. dr. S. Holm (University of Manchester, Universitetet i Oslo, Aalborg Universitet)
	prof. dr. C.J.H. Midden
	dr. J.H. Anderson (Universiteit Utrecht)
	dr. J.R.C. Ham

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Research for this thesis was made possible by the Netherlands Organization for Scientific Research (NWO) under grant number 313-99-160.

editors: Philip Brey, Anthonie Meijers and Sabine Roeser

ISBN: 978-90-386-4462-2
ISSN: 1574-941X

Copies of this publication may be ordered from the 4TU.Centre for Ethics and Technology, info@ethicsandtechnology.eu.
For more information, see http://www.ethcisandtechnology.eu.

Contents

1 Introduction 1

1.1. A first characterization of persuasive technology 3
1.2. Research questions and methods 6
1.3. Overview of the thesis 10
1.4. Bibliographical note 12

2 Ethics of persuasive technology: an overview 15

2.1. Introduction 15
2.2. A diagram for the ethics of persuasive technology 16
2.3. Ethical issues around use and development of persuasive technology. 21
2.4. Concluding remarks 41

3 Should we prefer nudging over rational persuasion? 43

3.1. Introduction 43
3.2. An account of rational persuasion 46
3.3. Rational persuasion and nudging contrasted: the account applied to nudging 54
3.4. Nudging and rational persuasion can be complementary 64
3.5. Conclusion 70

4 Persuasive technology redefined 71

4.1. Introduction 71
4.2. The Standard Definition and its Problems 72
4.3. Demarcating persuasive technology 74
4.4. Persuasive technology versus manipulative, coercive, and limiting technology 88
4.5. Conclusion 91

5 Buying more from an agent that looks like you. The ethics of similarity in persuasive technology. 95

5.1. Introduction 95
5.2. Similarity and 'computers as social actors' 97
5.3. Similarity, manipulation, and an ethics of communication. 100
5.4. Design guidelines for the use of similarity in PT 106
5.5. Objections 112
5.6. Concluding remarks 115

6 Scanlon on acceptable risk-imposition 117

6.1. Introduction 117
6.2. Scanlon's contractualism 120
6.3. Scanlon on risk-regulating principles 123
6.4. Interpreting Scanlon's appeal to intra-personal aggregation 127
6.5. Mitigating considerations 131
6.6. Concluding observations 134

7 The moral case for Intelligent Speed Adaptation 139

7.1. Introduction 139
7.2. Driving risks are morally problematic 141
7.3. Significant risk reduction through Intelligent Speed Adaptation 145
7.4. Harm prevention justifies ISA 147
7.5. Objection I: ISA introduces new safety risks 152
7.6. Objection II: ISA erodes the moral agency and responsibility of drivers 154
7.7. Objection III: Accepting ISA leads to a *Brave New World* society 157
7.8. Concluding remarks 159

8 The ethics of accident-algorithms for self-driving cars: An applied trolley problem? 161

8.1. Introduction 161
8.2. Programming Self-Driving Cars for How to React in the Event of Accidents 164
8.3. "The Trolley Problem"? 166
8.4. Two Very Different Decision-Making Situations 168
8.5. A Second Disanalogy: The Importance of Moral and Legal Responsibility 171
8.6. Stipulated facts and certainties vs. risks, probabilities, and uncertainties 174
8.7. Concluding Discussion 178

9 Conclusion 181

9.1. Conclusions to the research questions 181
9.2. Avenues for further study 184
9.3. Policy recommendations 185

Bibliography 187

Summary 207

Samenvatting 211

Curriculum Vitae 215

Simon Stevin (1548-1620) 219

Acknowledgments

What do you have that you did not receive?
1 Corinthians 4:7

I consider it as a wonderful privilege to have read, studied, discussed, taught, and written philosophy for more than four years, *as my profession*. Even though, more often than not, doing philosophical research was tough, at the most fundamental level it has been worthwhile and deeply satisfying. Here I wish to express my gratitude to all those who have been important to me along the way.

First of all, I wish to thank my daily supervisor, Andreas Spahn. Andreas, many thanks for your continuous optimism and confidence in me, which helped me every time to go on! My philosophical horizon has been widened by your 'big-picture tutorials' (during which you needed your whole whiteboard). You have a remarkable eye for ways in which one can (re)structure one's writings. However, what I enjoyed the most is your friendliness. It's been an honour to be your first PhD-student.

Thanks to Philip Nickel, my second supervisor. Philip, I was always looking forward to receiving yet another round of high quality feedback from you. You are a model for how one should do analytical philosophy, and you helped me a lot to improve my philosophical writing. I enjoyed your special sense of humor. Many people with your qualities are less modest. Thank you!

I also thank my promotor, Anthonie Meijers. Anthonie, I much appreciate the way you supervised my project over the years, your spirited manner of leading the supervisory meetings we had together with Andreas and Philip, and the way you helped me to integrate the feedback on my writing from you, Andreas, and Philip. Thank you for being well-organized, for your problem-solving attitude, and especially for always encouraging me.

I would like to thank the other members of my dissertation committee, Søren Holm, Cees Midden, Joel Anderson, and Jaap Ham for evaluating and providing feedback on my thesis.

Next our persuasive technology research project's members from the other disciplines: Thanks to Theo Hofman for introducing me to persuasive technology in the automotive domain. Thanks to Cees Midden, Jaap Ham, and Frank Verberne (my fellow PhD-student) for helping me to gain some understanding of the psychological perspective on humans using persuasive technology. Frank, I enjoyed working together with you on experiments, the organizing of the project's closing international workshop, and other practical matters regarding our project.

Part of the project was the policy workshop organized together with the Rathenau Institute in The Hague. I thank Jelte Timmer, Linda Kool, and Rinie van Est for a pleasant and fruitful cooperation. Rinie, thanks as well for our more informal discussions in Eindhoven.

The Eindhoven Philosophy & Ethics group has been a friendly and stimulating environment to work in, for which I would like to thank all my colleagues. Rianne, thank you for your help with all kinds of practical matters and for your special contribution to the nice atmosphere. Wybo, I admire and have benefited from your remarkable ability to both give penetrating feedback on my papers and to explain that they were going to be good enough: thanks. Joel, thanks for introducing me to teaching at the University. Martin, thanks for discussing the ethics of risk and many other topics, and for your advice, both solicited and unsolicited. Auke and Krist, thank you for wise counsel from the perspective of one that recently finished his PhD. Sven (Nyholm), writing papers together has been a revealing experience for me. It has clearly catalyzed my work in finishing my thesis during the period after the end of my contract. Thank you, also for your encouragement during that last stage!

To my fellow PhD-students, Christine, Marieke, Litska, Stefan, Sven (Diekman), Cletus, and Marjolein: I enjoyed our many chats about life as a PhD-student, including all its highs and lows. Cletus, thanks for being a calm and pleasant roommate. I will remember our attempts, under the guise of making cynical jokes, to analyse and get to grips with the phenomenon 'writers block'.

Outside of the academic world, I would first like to thank my current employer, Calvijn College in Goes, for generously permitting me to spend some of my hours dedicated to 'professionalization' to work on my thesis.

I thank my friends for bringing some necessary balance to my life as professional philosopher; I enjoyed playing games, hiking, cycling, discussing, and the like. Thanks for your involvement over the years. Machiel en Cors, your being my paranymph is just another way to celebrate our friendship.

To my family: thanks for being happy with me when I made the so much desired career switch, while you perhaps wondered what was so attractive about it. I appreciated your interest in my research, and discussing your objections against speed warnings and limiters was fun. Some of these objections have made it into this thesis, but ultimately have been found unconvincing...

Finally, dear Marian, thank you very much for letting me follow my heart, and for your continuous and loving support during the years of my PhD-project. And not in the least for your (sometimes critical) help with setting the right priorities. Together we are strong. Dear Simon, Maria, and Job, thank you bringing so much joy into my life! I encourage you to claim the lion's share of the extra time that becomes available now that I have finally finished 'my book'.

1 Introduction

On Saturday, May 7, 2016, Joshua Brown became the first driver of a self-driving car to be killed in a traffic accident. Both Brown and the Autopilot of his Tesla Model S failed to detect a left-turning truck-trailer combination, resulting in a fatal collision (The Tesla Team: 2016; Yadron & Tynan: 2016). In the Autopilot mode, the Tesla Model S performs several driving functions: advanced cruise control, lane keeping, and emergency braking, using information from a forward-looking camera, a radar sensor, and ultrasonic sensors. As a result, it is possible to let the car do all the driving and to keep one's hands off the steering wheel. However, Tesla explicitly intends and communicates its Autopilot as a safety *assistance* system, and not as a fully self-driving car. Therefore, drivers should always keep their hands on the steering wheel. In order to make sure drivers indeed keep their hands on the steering wheel and stay focused, the Tesla gives a visible and audible warning signal in case they fail to do so. These warning features are a piece of persuasive technology: technology that is explicitly designed to change people's behavior (Fogg: 2003, 1). Unfortunately, Tesla's persuasive technology failed to persuade Joshua Brown: during his fatal ride he ignored at least seven warnings. According to the investigation report of the National Transportation Safety Board, Brown received seven visible and six audible warnings, but had his hands off the steering wheel for 90% of the time (Fung: 2017; NTSB: 2017).[1]

This tragic accident clearly shows the need for ethical reflection on persuasive technology. The failure of the Tesla's persuasive technology led to, or at least contributed to the death of the driver, which raises several questions. To start with, can we regard this as a case of failed design, simply because of the magnitude of harm that results from the lack of persuasive success? Would, other

[1] According to Tesla, "What we know is that the vehicle was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the Model S. Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. The high ride height of the trailer combined with its positioning across the road and the extremely rare circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S"(The Tesla Team: 2016).

persuasive strategies, besides or instead of audible and visual warnings, have been more successful? Relatedly, if the inability to persuade the driver could lead to his death, would it not be better to rely on technology that *forces* drivers to keep their hands on the steering wheel? Whereas for some purposes it is clear that users should be granted an autonomous choice whether or not to perform the behaviour intended by the persuasive technology, in case of the Tesla the driver is perhaps being granted too much autonomy? In this regard, it is interesting that one of the Tesla software updates subsequent to the accident implemented a strategy that is closer to coercion: when a driver repeatedly ignores the Tesla's warning signals, the Autopilot function may become unavailable until the next drive (NHTSA: 2016).

The Tesla accident points to the different responsibilities of the various parties involved. The Tesla engineers should see to it that they design a car in accordance with the relevant safety standards. This includes ensuring that drivers, or at least average and reasonable drivers of good will, use the car properly. Of course, drivers have the responsibility to use the car in ways they are instructed to. Tesla's sales people need to make sure that they raise factually correct and realistic driver expectations with regard to the capabilities and proper use of the autopilot. Some critics, however, hold that Tesla, despite its urging drivers to keep their hands on the wheel, should stop advertising with the term 'Autopilot', because it misleadingly suggests that drivers need not pay attention to their driving tasks (Reuters: 2016). Other involved parties are the various national traffic safety boards that issue regulations, monitor compliance, and investigate accidents. We see how the interplay between the various involved parties and their different role responsibilities make it not so easy to tell which party is preeminently to blame for the fatal accident. And consequently, it is not immediately clear who should do what in order to prevent similar accidents in the future.

At a more general level, the Tesla case illustrates how persuasive technology can be involved in various value conflicts and value trade-offs. Cars are more and more equipped with advanced driver assistant systems intended to make traffic safer, more sustainable, and more efficient. Safety, sustainability and mobility are all social values held in high regard, but at the same time, they may be only achievable at the cost of reduced driver autonomy, privacy, and perhaps other values. Driver autonomy may be at stake if the driver's freedom to drive in ways he sees fit is diminished, and the vast amount of data recorded by the assistant systems is an imminent privacy threat. One task of an ethics of persuasive technology, taken up in this thesis, is to analyze such value conflicts, to find out how exactly they arise, how the conflict could be eased by adapting the design, and which value trade-offs are acceptable. Before setting out the research questions of this thesis in a more systematic way, I will now first briefly introduce the concept of persuasive technology and situate it relative to other (technological) means of influence.

1.1. A first characterization of persuasive technology

As already stated, persuasive technologies are technologies explicitly designed to change human attitudes and behaviour. In addition, it is definitive of persuasive technology that its methods do not rely on coercion or deception, because persuasion implies voluntary change. (Fogg: 2003; IJsselsteijn, Kort, Midden, Eggen, & Hoven: 2006). Consequently, persuasive technologies by definition place its users in a position of ultimate control over their attitudes and behaviours. It is instructive to locate persuasive technology on a control continuum analogous to the various ones defined for automation technology (SAE International: 2014). On the left end of these scales, all control is allocated to the human driver who performs all driving tasks. On the right end, all control is allocated to the car: it is fully autonomous and performs all the driving tasks. Moving through the intermediate levels, the car increasingly takes over driving tasks, starting with warning functions. On such kind of a control continuum, persuasive technologies are close to the left end. A persuasive technology may provide behavioural suggestions, reasons for the behaviour, information needed to perform some behaviour, social support to overcome a lack of will-power, and so on. In addition, it sometimes also provides task support, such as the laborious calculation of fuel consumption by eco-feedback technology, which eases performance of the persuasive technology's target behaviour. But, persuasive technologies do not completely take over human actions, and they also leave users the choice whether or not to perform the intended behaviour. Persuasive technologies by definition do not force, nor deceive or manipulate its users. The largest share of control is allocated to the user.

Persuasive technologies form a rapidly growing class of technologies. As is widely recognized, many societal problems have a large behavioural component, and this is where politicians and others see the potential for persuasive technologies as a means to help in solving these problems. We need to significantly reduce our energy use and CO_2 emission; obesity is a growing problem; our medical and mental health care systems are taking increasingly larges shares of government budgets; and so on. For many societal problems, persuasive technologies could help to change people's behaviours. For example, eco-feedback systems persuade car drivers to use less fuel by means of images of green leaves that appear (Honda: 2008). Special websites or health apps try to motivate and support people to exercise regularly and eat healthfully, or to quit smoking . They do so by means of giving tailor-made suggestions for healthy behaviour, giving social praise, or, as already noted, giving task support, e.g. counting calories. The BabyThinkItOver infant simulator persuades teens not to become pregnant, by providing them the experience of baby caring (Fogg: 2003, 78–79). Internet based coaching programs with various persuasive features are being developed to support persons with specific needs, such as autism (Mintz & Aagaard: 2012; Odom et al.: 2015). We see that various persuasive technologies have been developed with the aim to foster particular social values.

But in addition, commercial use of persuasive technologies is rapidly increasing and has a large potential. For example, web retailers try to persuade customers to buy products, by giving personalized suggestions for what they might like, based on previous online behaviour. Increasingly, artificial agents equipped with various social features are employed as digital assistants and sales-agents (Qiu & Benbasat: 2009).

The phrase 'persuasive technology' is less well-known than the underlying concept of technology designed to influence people. The phrase was coined by B.J. Fogg, author of the book "Persuasive Technology. Using Computers to Change What We Think and Do" (2003). This book soon became a classic in the emerging field of the study and design of persuasive technology. That this field is indeed young is illustrated by the fact that the first international conference on persuasive technology was held in 2006 in Eindhoven, The Netherlands (IJsselsteijn, Kort, Midden, Eggen, & Hoven: 2006). From 2006 onwards, there has been an annual conference that attracts several hundred (social) psychologists, designers, computer scientists, a few ethicists, and a few scholars from other disciplines such as neurology. Persuasive technology is clearly an interdisciplinary topic.

The concept of persuasive technologies has similarities to and some overlap with other concepts of technologies with an aim to influence behaviour. First we can mention e-coaching, which refers to online coaching programs, often with specific goals to learn new behaviour that is conducive to better health and well-being (Kool, Timmer, & Est: 2015b; Warner: 2012). The inclusion of features to motivate users, for social support between users, etc. gives e-coaching much resemblance to persuasive technology. Second, there is the concept of Ambient Intelligence. Ambient Intelligence refers to “sensitive, adaptive electronic environments that respond to the actions of persons and objects” (Aarts & Wichert: 2009). One can think of elderly people wearing sensors that register a fall and warn a care-giver. This technology is often also designed to bring about specific user behaviour, thereby qualifying as persuasive technology (Kaptein, Markopoulos, de Ruyter, & Aarts: 2010). Third, there is the hotly debated approach of nudging. Nudging refers to the use of scientific knowledge of the way humans choose and decide in order to gently steer people to choose and behave in ways that would benefit them. Nudging is all about deliberately rearranging the decision-making situation in such a way that desirable choices become more likely. Think of placing healthy food at eye-height, making saving for pensions the default option instead of not saving, and the like. Importantly, nudge proponents stress that nudging does not forbid any option, thus it would be no threat to human freedom and autonomy (Sunstein & Thaler: 2003; Thaler & Sunstein: 2009). This seems analogous to Fogg’s insistence that persuasive technology implies voluntary change. We will see that in fact, the threat of manipulation looms large for nudging; in case of persuasive technology we need to be careful as well. In any case, to the extent that nudging depends on technology as a means, it often resembles persuasive technology.

Given these resemblances, much of the ethical reflection on persuasive technology in this theses is relevant for these other influence concepts as well. Nevertheless, the discussion in the chapters below primarily focusses on persuasive technology as defined and researched under the heading or label of persuasive technology. This is done for two reasons. First, I explicitly intend my thesis to be useful, first of all, to the community of persuasive technology scholars, designers, and users. The hope is that connecting to the literature and discussions in this field helps to achieve that aim. Second, focusing on persuasive technology as understood and studied in the distinct field of persuasive technology is a useful way to focus and delineate this thesis’ research. But again, much of this thesis will be relevant to e-coaching, ambient intelligence, and nudging, and it will often not be very difficult to apply the findings of this thesis to these other influence concepts. In the case of nudging, this research does engage rather explicitly with the concept, as a way to get clear on the what exactly persuasive technology is.

1.2. Research questions and methods

The research questions of this thesis can be summarized under the heading of one overarching question: how can persuasive technologies be designed, implemented and used in an ethically responsible way?

In order to answer this question, careful attention will be paid, throughout this thesis, to how control is distributed between users and the persuasive technology. Thus, in terms of the thesis' subtitle, a central focus is on the allocation of control. This is useful and illuminating, because the locus of control is ethically crucial in several ways. Firstly, it will turn out to be essential that users of persuasive technology maintain substantial control over their mental states and behaviour. If they do not enjoy such control, there likely will be manipulation or coercion going on. In other words, their autonomy is not given due respect, because personal autonomy involves exercising some sort of control over oneself. Therefore, designers face the challenge to both successfully bring users to change their behaviour, and to do so in a non- controlling manner. Careful analyses of various strategies and means of persuasion will be carried out at several places in this thesis in order to determine whether or not this is the case. Secondly, and relatedly, for some social values, such as traffic safety, it could be argued that user control over technology should be limited. Traffic safety might be regarded as such an important end that the user has to give up some autonomy by accepting Advisory Intelligent Speed Adaptation mandated by the government, or even speed locks. In case of self-driving cars, the topic of Chapter 8, control is even fully transferred to the technology.

Of course, the overall research question is too extensive to be answered in depth with regard to all its aspects. Chapter 2 will give a summarizing treatment of the majority of ethical issues involved in design, implementation, and use of persuasive technologies. Four main ways can be distinguished in which the remainder of the thesis is focused and goes into more depth, corresponding to four main research questions. The first question connects to the requirement that persuasive technology relies on voluntary change of the user's attitudes and behaviour. Whereas the concept of 'voluntariness' has a rather clear intuitive sense, it turns out not to be so easy to operationalize this concept with regard to persuasive technology. The present research rephrases the issue of 'voluntary change' in terms of 'substantial control of users over their attitudes (and other mental states) and behaviours'. The question then becomes under what conditions users exercise such control. In Chapter 3, a comparison is made between rational persuasion and nudging. From this comparison, conditions are derived that must obtain for humans to be in control over their attitudes and behaviour. In Chapter 4, these control conditions are adapted to and specified for the case of persuasion by means of technology, that is, for persuasive technology. In Chapter 5, a case study is conducted to investigate the human user's measure of control over her behaviour and mental states with regard to one prominent persuasive strategy. This strategy concerns the use of a specific form of so-called social influence, viz. 'similarity influence' (i.e. influence based on perceived similarity between the persuasive technology and oneself as user).

The second focus of this thesis immediately relates to the first. It concerns the definition or characterization of persuasive technology. The definition given above is less clear than it might seem on first glance. It mainly defines the 'persuasion' part of persuasive technology in terms of what it is *not*: not coercion, and not deception. But a more positive statement of the underlying mechanisms of persuasion in persuasive technology was absent from the literature. Consequently, in the field of study and design, there is some confusion about what does and does not qualify as persuasive technology, and why. Sometimes pieces of coercive or of manipulative technology are presented as persuasive technology. Also, sometimes the mere fact that some technology influences human behaviour is taken as reason to call it persuasive technology. The question, then, becomes how to define persuasive technology in a way that clearly characterizes what it is, and how it is different from other types of technological influence. In Chapter 4, building on the discussion of control in Chapter 3, a redefinition of persuasive technology is developed that aims to solve the problems mentioned. This redefinition better enables us to distinguish persuasive technology from other types of influencing technologies.

A third main question of the research of this thesis is how to morally evaluate the use of social influence strategies in the design of persuasive technology. 'Persuasion' as research topic belongs to the field of social psychology because persuasion as such is a form of social influence. It is very clear that the majority of persuasive technology design goes beyond exchanging factual information with users and employs several types of social influence to enhance persuasion (Torning & Oinas-Kukkonen: 2009). However, our moral evaluation could be significantly different for a given means of social influence in human-human interaction versus in persuasive technology-human interaction. What is natural, unintentional and morally unproblematic in human-human interaction, may easily become a form of morally wrong manipulation when employed in the design of persuasive technology. Many examples of social influence are considered in Chapters 2, 3, and 4, in order to gain an understanding as to how different types of social influence affect human control over attitude and behaviours. In Chapter 5, the case study explicitly addresses similarity influence in order to investigate what exactly changes from an ethical point of view when social influence is transferred from its natural context of interacting humans to the design of a persuasive technology's communication with users.

These three main questions, addressed in Chapters 2, 3, 4, and 5, all focus on the means of persuasion: what psychological mechanisms are involved in the persuasive technology's operation and how do these affect user control? In the second part of the thesis (Chapters 6, 7 and 8), there is a shift towards the ends of persuasion, and more specifically to how the technological means are justified so as to change user behaviour, relative to different ends. Mobility, and in particular auto mobility serves as the context of this discussion. In mobility, there are prominent social values: safety, mobility, and sustainability. Mobility is an interesting domain to analyse from an ethical perspective, because many of the efforts of designers are aimed at making traffic cleaner, more efficient, and safer. Technologies such as eco-feedback, lane-departure warning, speed advice and speed locks, and ultimately the self-driving car are claimed to foster these values. However, they may be in tension with the autonomy of drivers, especially automotive technologies that limit driver control and allocate control to the car. In order to assess whether the government may legitimately mandate these automotive technologies, we need a convincing ethical evaluation of current driving risks, and of the general safety level of traffic. One issue is when to mandate persuasive technologies, and when to take recourse to stronger measures, such as robot technology that controls the car with respect to one or more driving tasks.

Accordingly, the fourth and final focus of this thesis concerns the relation between persuasive technology and acceptable risk. A plausible framework is needed to determine which risks are unacceptable, and could therefore justify limiting citizen autonomy by obliging risk reduction through the use of persuasive technology.[2] Chapter 6 addresses the question of when and why it is acceptable to impose risk on others, taking the perspective of Scanlon's contractualism. This is a difficult question, and one that has been an increasing focus of ethics research (Altham: 1983; Hayenhjelm & Wolff: 2012). Scanlon's contractualism is an especially promising approach, since it seems to provide a reasoned middle ground between consequentialist accounts, which are seen by critics as overly permissive, and rights-based or Kantian approaches, which are seen by other critics as overly restrictive. In Chapters 7 and 8, conditions of acceptable risk imposition play a justificatory role in the discussion of Intelligent Speed Adaptation (a piece of persuasive technology), and autonomous vehicles (a piece of automation technology).

After setting out the research questions, it is helpful to explain the methods used in this thesis. It will be clearly visible that the present research was conducted as part of an interdisciplinary project. Mechanical engineers, social psychologists specializing in human-technology interaction, and philosophers all cooperated in this NWO-funded research project.[3] The research was focused on persuasive technology in the automotive domain, for example eco-feedback technology.

As a result, the traditional method of conceptual analysis used in this thesis (Margolis & Laurence: 2015) has been substantially empirically informed by the literatures on social psychology, human-technology interaction, automotive technology, and traffic safety. Conceptual analysis is employed to analyze various concepts, most importantly persuasion, rational persuasion, and persuasive technology. However, this is not a matter of armchair philosophizing, in which concepts are analyzed into their constituent parts by means of reflection on one's intuitions regarding the concepts. Rather, social psychological theorizing on persuasion, philosophical accounts of argumentation, argumentation theory, and findings of the human-technology interaction literature are integrated with conceptual analysis. It is also worth noting that the interdisciplinary setting of the present research also led to two case studies. Chapter 5 investigates similarity-based influence, because this type of influence was extensively studied in the

[2] In the second place, the use of persuasive technology may also itself generate risk, for example risks connected to a failure to persuade users to perform the target behavior (as in the Tesla case described above).

[3] NWO stands for The Netherlands Organisation for Scientific Research, https://www.nwo-mvi.nl/project/persuasive-technology-allocation-control-and-social-values

social psychology part of the research project (Verberne: 2015). Chapter 7 investigates a piece of automotive technology, namely Intelligent Speed Adaptation, which warns speeding drivers or makes speeding technologically impossible.

Like any research project in applied ethics, the present research has had to deal with moral pluralism and the substantial disagreement between ethical theories and frameworks. Rather than committing itself firmly to one of the competing approaches to ethics, in the first half (Chapters 2 until 5) this thesis aims to build as much as possible on general ethical principles and values shared widely and to a large degree. This is particularly important in the case of moral values such as autonomy, privacy, sustainability, and the like. These values are widely recognized as important by different ethical theories, but often for different reasons. The same point holds for principles such as 'do not manipulate'. According to deontological or Kantian ethical theories, this principle directly derives from the supreme value of autonomy. According to utilitarianism, this principle is conducive to higher overall utility, because people themselves know best how to lead their own lives. By focusing on shared moral values and principles, this thesis aims to provide results acceptable to many readers. But of course, this is not an easy way out of moral pluralism. For even though these values and principles are shared, there may still be significant disagreement as to their priority and relative weight, and their applicability. In the second half of the thesis, during the discussion of acceptable risk, more weight is given to considerations of distributive justice. Such considerations are usually thought to be more at home in contractualist and deontological ethical frameworks.

1.3. Overview of the thesis

Although the structure of this thesis is already visible in the exposition of the main research questions, it will be helpful to give a more detailed overview of the individual chapters and to indicate how they are interrelated.

Chapter 2 summarizes and extends the existing literature on the ethics of persuasive technology. Its main contribution to this literature is twofold. First, it disambiguates the concept of 'outcome' into the behaviour that results from user interaction with the persuasive technology on the one hand, and how that behaviour has an impact on central social values, e.g. health, sustainability, and the like, on the other hand. Second, instead of looking at designers alone, the interaction between designers and deployers of persuasive technology is shown to confer additional insights. Several ethical issues of a more emerging nature are discussed as well, such as the risk that extensive use of persuasive technology leads to deskilling and weakened capacities for self-regulation.

Chapter 3 provides a broader background to the philosophy and ethics of persuasive technology by contrasting 'nudging' (a concept related to and partly overlapping with persuasive technology) with rational persuasion. The psychological processes underlying persuasion are studied as a way of determining which 'non-argumentative means of persuasion' grant the user of persuasive technology substantial control over her mental states and behaviour. Three guidelines are proposed for guaranteeing such control, and are subsequently applied to contrast nudging with rational persuasion. It is concluded that under certain conditions, nudging and rational persuasion are complementary, not contradictory.

Chapter 4 builds on the results of Chapter 3 in order to redefine persuasive technology, as technology that influences by means of communicating with users in a way that grants them substantial control. The field of study and design of persuasive technology is unclear about the precise meaning and boundaries of the concept of persuasive technology. In order to adequately deal with this situation, this thesis has taken a detour and first discussed existing literature and situated 'persuasive technology' in a broader context. The results of this exercise have proven indispensable to the project of redefining persuasive technology. The improved definition allows to better distinguish persuasive technology from manipulation and coercion.

Chapter 5, as indicated already, zooms in on the use of one particular type of 'non-argumentative means of persuasion', viz., similarity-based influence. This concerns ways in which the persuasive technology, often in the form of artificial social agents, is made similar to its users, for example by mimicking the user's speech pattern and head-movement. It is argued that the use of such influence can easily turn into manipulation and therefore raises genuine concerns that users are insufficiently in control of their own mental states and behaviour. Three guidelines are developed for responsible use of similarity-based influence. As already noted above, the next chapters focus on the ends of persuasive technology, as distinct from the means discussed in previous chapters. In mobility, the crucial end or value is *safe* mobility, or mobility without unacceptable risk.

Chapter 6 deals with the ethics of risk by way of preparation for chapters seven and eight. It contains a discussion of how Scanlon's contractualism deals with the question of acceptable risk-imposition. It argues that Scanlon's commitment to avoid inter-personal aggregation leads to very stringent restrictions on morally acceptable risk imposition. Whether these restrictions are plausible or not is a matter of debate, but three considerations that follow from Scanlonian contractualism deserve further reflection and practical application. First, it is likely that our society should do more by way of taking precaution and reducing risk-imposition. Second, in doing so, it should focus on the aggregated amounts of risk faced by individuals during their lifetimes. Third, a system of mutual equal risk-imposition will only be fair if it includes all the earth's citizens, including future generations (although it is not immediately clear how these could be represented in contractualist reasoning).

Chapter 7 argues for a positive moral case for mandatory Intelligent Speed Adaption, based on its potential to significantly reduce driving risks and the resulting harms. In the course of doing so, it contrasts a persuasive version of intelligent speed adaption (which warns drivers when speeding) with a limiting form (which makes speeding technologically impossible). The chapter reviews all possible objections and concludes that governments should mandate the use by all drivers of the strongest form of Intelligent Speed Adaptation, i.e., the limiting version.

Wheras Chapter 7 features technologies that are already high on the control continuum, Chapter 8 deals with self-driving cars, to which (nearly) all control is allocated. The chapter argues against conceiving the problem of programming autonomous cars for situations of unavoidable accident as an applied trolley problem. It does so on the basis of three major disanalogies. One of these disanalogies concerns the fact that in trolley problems decisions can be made on the basis of full and certain knowledge of certain outcomes, whereas in case of programming self-driving cars, we have to deal with partial and uncertain knowledge of probabilistic outcomes. In short, this chapter illustrates that a well-developed ethics of acceptable risk-impostion is indispensable for deciding on how control should be allocated.

Chapter 9, finally, summarizes the main conclusions of the thesis and points out several avenues for further study.

1.4. Bibliographical note

Chapter 4 partly draws on a previous publication, and Chapters 7 and 8 are very slightly modified versions of previously published papers.

Chapter 4

Ideas presented in Chapter 4, in particular in section 4.4, are drawn from: Smids, J. (2012). The Voluntariness of Persuasive Technology. In M. Bang & E. L. Ragnemalm (Eds.), *Persuasive Technology. Design for Health and Safety* (pp. 123–132). Springer Berlin Heidelberg.

Chapter 7

Smids, J. (accepted for publication). The Moral Case for Intelligent Speed Adaptation. *Journal of Applied Philosophy,* forthcoming.

Chapter 8

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289.

This chapter was written together with Sven Nyholm (first author), and I thank him for permission to include it in this thesis. I wrote section 8.2 and significant parts of section 8.6 (I extended our scenario of the elderly pedestrian and wrote parts on the ethics of risk). We extensively commented on each other's sections and had several rounds of discussion and editing. The chapter is included in the present dissertation, because self-driving cars nicely illustrate ethical aspects of automotive technology at the right end of the control continuum, which will probably follow up current persuasive technology in the domain of mobility.

2 Ethics of persuasive technology: an overview

2.1. Introduction

In this chapter, I will give an overview of the ethics of persuasive technology. Building on the literature, I will present a diagram that depicts ethically relevant agents, aspects and values surrounding persuasive technology (PT). This diagram serves as an analytical tool that should assist scientists, designers and deployers, regulators, users, and society in thinking about the moral aspects of persuasive technology. Accordingly, I use it in this chapter to give a relatively comprehensive treatment of ethical issues that will arise around PT rather than an in-depth treatment of one particular aspect, such as that given in Chapter 5.

This chapter does not take one ethical theory as its normative starting point but, instead, a set of moral considerations that are widely supported and recognized as relevant for designing technology. One can think here of principles of non-manipulation (derived from respect for autonomy), fairness and justice, respect for privacy, and other principles that have multiple sources of support.[4]

The structure of this chapter is as follows. In the next section (2.2), I will first discuss an existing diagram (Berdichevsky & Neuenschwander: 1999). Next, I will modify and expand that diagram into a diagram that depicts the most relevant agents and aspects surrounding the design and use of PTs, including some of their most important relations. In section 2.3, I will discuss the main ethical questions and issues related to persuasive technology. I will identify these questions by analysing how the diagram's agents are related both to each other and to the PT. This analysis will be enriched by considering the agents' roles, expertise, responsibilities, normative justifications, and values. I will summarize the major ethical issues in section 2.3.5. In section 2.4, I will make a few concluding remarks.

[4] Of course, ethical theories will differ as to why these principles are important, and also on their relative importance, and on their more precise meaning. However, this chapter applies the principles at such a level of generality that these debates can be side-stepped.

2.2. A diagram for the ethics of persuasive technology

In 1999, Berdichevsky and Neuenschwander published their pioneering and influential article '*Toward an Ethics of Persuasive Technology*' (Berdichevsky & Neuenschwander: 1999). In that article, they provide a helpful diagram that relates persuasive technology to two crucial agents: designers and users who are to be persuaded (Figure 2.1).

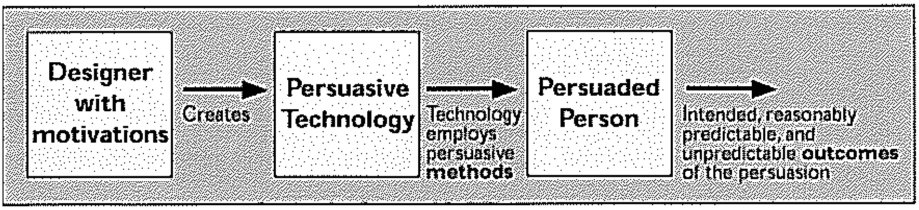
Designer with motivations
Creates
Persuasive Technology
Technology employs persuasive methods
Persuaded Person
Intended, reasonably predictable, and unpredictable outcomes of the persuasion

Figure 2.1 Berdichevsky and Neuenschwander's diagram of ethically relevant agents and relations (op cit., 54)

Using the diagram, the authors introduce the conceptually and morally important distinction among designer motivations, persuasive intent, persuasive methods, and persuasive outcomes. Whereas designers have their professional and personal motivations, the persuasive intent is attributed to the persuasive technology. This refers to what the PT is supposed to persuade a person to do or think. Berdichevsky and Neuenschwander explain their diagram, which they call a framework,[5] with the following example (slightly adapted for our purposes). Suppose three designers each build a PT with the persuasive intent to get some person to eat more fruits and vegetables. Their personal motivations may differ: the first designer may be motivated to increase the person's health, the second may be motivated to increase the profits of farms that grow fruits and vegetables, and the third may be motivated "by a secret hope the [person] will eat a bad fruit and become sick to the stomach" (op cit., 55). In this example, the persuasive intent is the same, but the designer motivations behind the intent differ in an ethically important manner. The outcome of the technological persuasion is defined by the authors as "what the persuaded person is persuaded to do or

[5] Berdichevsky and Neuenschwander call it a framework, but depending on how we understand that term, it might be more apt to speak of a 'conceptual diagram'. An ethical framework is commonly understood as also including ethical principles that provide normative guidance for purposes of evaluation.

think". Here, this is the consumption of fruits and/or vegetables, possibly as a result of newly acquired positive attitudes.

Although Berdichevsky and Neuenschwander's diagram is a useful starting point, it has a serious problem and needs expansion. The problem is rooted in both the ambiguity of the concept 'outcome' and the manner in which this ambiguity is reflected in Berdichevsky and Neuenschwander's use of the concept. They define 'outcome' as referring to attitudinal and/or behavioural change. At some points in their paper, however, 'outcome' also refers to the *subsequent effects* of these changes.[6] This easily leads to confusion and even in cases when it does not, we will achieve additional conceptual clarity and analytical tools by carefully distinguishing these two senses of 'outcome'. This is important because most often, we are less interested in behaviour *as such* than we are in behaviour's impact on the world (including on the users themselves), explicated and evaluated in terms of certain important ends and values. Thus, to follow the present example, we are interested in the impact of eating more fruits and vegetables on a person's health and well-being, on her longevity, on the flourishing of the local economy, and so on. The promotion of these values—health, personal well-being, and wealth—is the goal or the end for which the designers developed the example's PT. I will refer to these values interchangeably as the PT's 'target value(s)' and 'end(s)'.[7]

To avoid the ambiguity connected with 'outcome', I will distinguish between the behavioural changes induced by the PT and the subsequent impact of those behaviours on the world. To fully grasp a PT's impact, it is useful to introduce two further, more fine-grained distinctions. First, we note that in addition to or even instead of the target behaviour,[8] a PT may lead to unintended behaviours. Second, in addition to the values that the PT should foster, other values are usually affected, whether positively or negatively.

[6] This is evident from their example of the user dying from an allergic reaction to digesting a kumquat. They refer to this tragic event as the unintended outcome, while according to their terminology this is rather a consequence of the outcome "eating a kumquat". This ambiguity is visible throughout their paper and also in Fogg's discussion of the ethical importance of outcomes, which builds on Berdichevsky and Neuenschwander (Fogg: 2003, 227–230).

[7] Sometimes the PT's end, or target value, is part of a chain of ends that ultimately lead to a final end.

[8] Note that it would be more precise to speak of the PT's target mental states, since it is *via* influencing mental states that the PT influences subsequent behavior. For present purposes, the 'target behavior' is meant to include these preceding mental states.

The diagram below results from expanding and modifying Berdichevsky and Neuenschwander's work to include the distinctions just introduced.

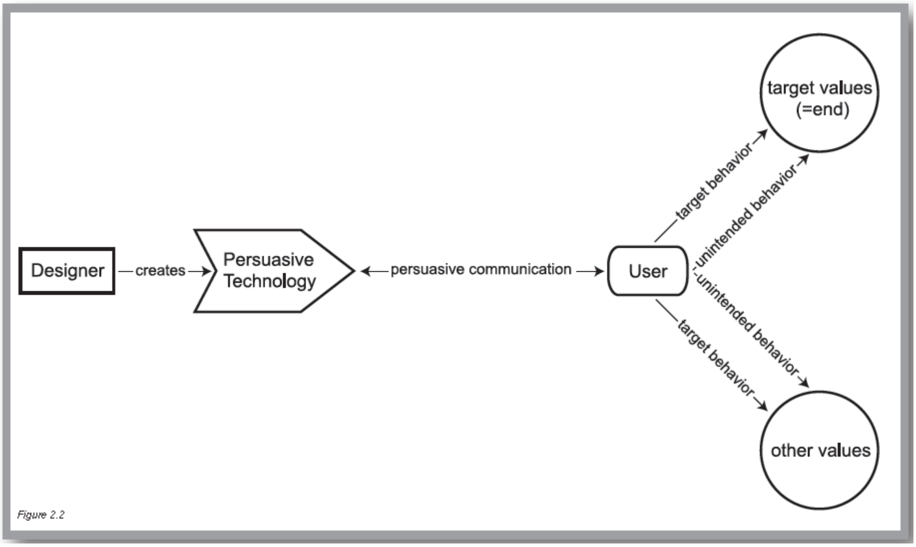

Figure 2.2 Diagram of ethically relevant agents and relations, expanded and modified from (Berdichevsky & Neuenschwander: 1999).

The four arrows between 'user' and 'values' represent four possible causal pathways along which a PT may have an impact on the world. If the user performs the target behaviour, this may—and often does—affect both the target values and other values. The same is true for unintended behaviour. In addition, each impact along one of these four pathways may be either positive or negative. This leads to as many as eight logically distinct possibilities that must be considered when analysing a PT's influence. In practice, of course, not all of these possibilities will be either equally important or likely to occur. To follow the present example, when the user engages in the target behaviour—eating the kumquat—the PT's target values (user health and well-being, farms' economic profit) are fostered. However, Berdichevsky and Neuenschwander also envision the much more remote possibility that the user will die from an allergic reaction to digesting a kumquat (op cit., 55). In this case, the PT successfully persuades the user to perform the target behaviour but utterly fails to foster user health.

The target behaviour may also affect—either negatively or positively—values *other* than the target values. Accordingly, if the fruits are very expensive, low-income households may encounter financial problems that negatively affect both their income and their opportunities in life. In this case, the PT has a mixed impact on the user's overall well-being. Thus, it becomes clear that a negative effect may result from the target behaviour itself, not only from unintended behaviour (which initially seems more likely to occur). This is one benefit of the sharp distinction between behaviour and its consequences that is made in the diagram. In the same spirit, the diagram invites us to consider the possibility that unintended behaviour may also have positive effects.

In addition to the distinction between behaviour and its consequences, it is worthwhile to make two further modifications to Berdichevsky and Neuenschwander's diagram. The first involves substituting 'user' for 'persuaded person'. It is more apropos to speak about users, given the fact that people often actively use the PT instead of merely being the subject of its persuasive attempt. The second modification is closely related: the term 'persuasive communication' will be used instead of 'persuasive methods'. As will be argued in Chapter 4, it is a defining characteristic of PTs that persuasive methods first and foremost involve communication, albeit of a simple form. Presenting users with certain types of information—often some kind of feedback—is always part of the persuasive attempt. Since users often actively engage with the PT—e.g., by providing requested information or by changing certain settings (which involves the input of information)—'persuasive communication' is depicted as a bidirectional relation. Once we realize that PTs communicate with users, we can draw from the ethics of communication to develop an ethics of PT (Spahn: 2011). I will do so in Chapter 5.

After being modified in the ways just described, the diagram remains incomplete in some significant ways and needs extensions. A crucial yet absent agent concerns the 'deployer'. Often, designers will develop a new persuasive technology because a party, who I will call the 'deployer', has commissioned it. It is the deployer who attempts to 'sell' the PT to specific groups of users, aiming to influence their behaviour. As we will see in the next section, designers and deployers will have both different motivations and different moral responsibilities regarding PT, which is why it is useful to distinguish them.[9] For example,

[9] Of course, this still involves gross simplifications. First, it would be more accurate replace "designer" with the concept of a "Research and Development Network" that can be described as "that part of the innovation chain that builds upon the results of fundamental research and is followed by product development" (Doorn: 2011). Second, the PT developed will also have to be manufactured or produced. Because generally this production phase doesn't raise ethical issues that are distinct for PT, it is not considered in this chapter. Third, the PT once developed

PTs usually generate a great deal of user data that could be accessible to deployers, giving rise to the issue of privacy.

Another useful addition to the diagram concerns the category of what I will call 'context'. The purpose of this category is both to avoid forgetting and to make room for all other relevant parties, most importantly, governments, regulators, and people closely related to users. This context, for example, health care, education, traffic and mobility, will often shape the agents and their mutual relations. For example, in a society with little government regulation of PTs, less-educated users who have only a few friends may be vulnerable to abuse—specifically, manipulation—by commercial deployers of PTs. In the same society, users who are not vulnerable in those ways may instead profit from the PT, seeing through manipulative attempts to influence them.

Finally, arrows from the 'target values' both to users and to deployers indicate that the wish to foster specific target values motivates not only deployers but often also the users themselves. In the latter case, the PT can be seen as a self-help tool for users to achieve some values they find important. Figure 2.3 below shows the diagram that includes these further additions. In the next section, I will show how this diagram is helpful for discovering and organizing ethical issues arising around the use and development of PTs.

and produced will often be marketed. Finally, preceding the research and development phase, there will be fundamental research that serves as input. I will make some brief remarks on ethical obligations of scientists working on psychological influence strategies in 2.3.4 below.

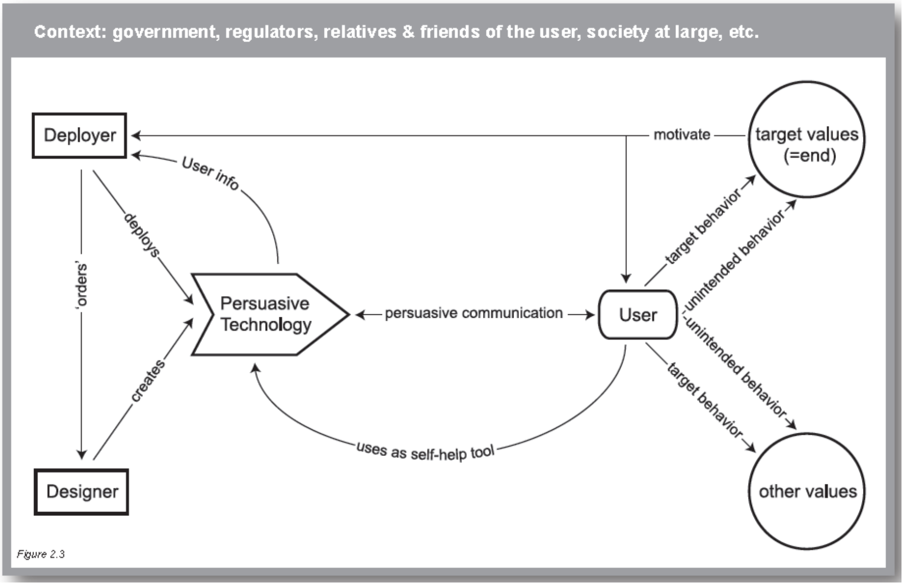

Figure 2.3 Final diagram of ethically relevant agents and relations concerning persuasive technology.

2.3. Ethical issues around use and development of persuasive technology

In this section, I will use the diagram developed in the previous section to introduce and discuss the main ethical questions and issues that arise in connection to the development and use of PT. We will arrive at a relatively comprehensive overview of the ethics of PT by thinking through *how* each of the diagrams' agents is related both to the other agents and to the PT. This exercise includes investigating which values and interests are crucial to each agent, which justified normative expectations each agent holds, which specific expertise or knowledge each agent has, and which responsibilities can be plausibly attributed to each agent.

The discussion of ethical issues will be organized by dividing the diagram into four clusters of agents. Firstly, I will look carefully at the centre of the diagram: the interaction between the PT and the users. Secondly, I will deepen the previous section's discussion of how user behaviour, both intended and unintended, connects both to the target values and to other values. Thirdly, I will discuss the role of deployers and designers, focusing on their responsibilities with regard to the ethical issues identified at that stage. Fourthly, I will investigate what additional insight we gain by considering the broader context in which PTs are used and developed. This will include some questions that transcend the level of individual agents, such as whether the identified agents' individual responsibilities are sufficient to safeguard ethically sound PT.

2.3.1. Persuasive technology – user interaction

When reflecting on the communicative interaction between persuasive technology and users, we arrive at PT's single most acknowledged ethical issue: its impact on the user's personal autonomy (Anderson & Kamphorst: 2014; Schermer: 2007; Spahn: 2011; Verbeek: 2006). This is because PTs aim to influence users, and any attempt to influence people carries the risk of manipulating them or otherwise failing to respect their autonomy. According to the *communis opinio* of PT scholars, it is a defining aspect of PTs that they "rely on voluntary change" (Fogg: 2003, 1,15,16; cf. IJsselsteijn et al.: 2006; Oinas-Kukkonen: 2010). Technological methods of persuasion should not obstruct or threaten personal self-government by failing to satisfy that voluntariness condition. Below, I will explain several ways in which user autonomy is vulnerable to ill-designed or misused PTs and discuss how to address each of those vulnerabilities. That is, I will propose several guidelines that help safeguard ethically sound interaction between a PT and its users. First, however, I will expand on the meaning and value of autonomy.

Broadly speaking, an autonomous person is the author of her own actions. She governs her life in light of her own beliefs, values, aims, personal commitments, life plans, and the like.[10] The autonomous person is not the mere product of her desires, external forces, opinions of others, etc.: she is also capable of sometimes standing back and asking herself how she wants to relate both to that which moves her to act how she does and to the circumstances in which she finds herself. Although as a human being, the autonomous person is socially embedded and need not be able to subsist independently of others, she can critically reflect on what others believe and want her to do and act on her own judgment (Christman: 2011, 2015; Dworkin: 1988; Oshana: 2006).

So far, I have been giving the broad idea of what is commonly called the '*condition* of autonomy', the state of actually governing one's life (Feinberg: 1989). Persons can be autonomous to significantly differing degrees depending on how

[10] For purposes of this chapter, it suffices to identify the commonly recognized core idea of autonomy, and to explain why it is important, without going into the extensive literature on the concept of autonomy.

they exercise their '*capacity* for autonomy'. Having a general capacity for leading an autonomous life means having several competencies, most notably for self-reflection, rational thought, and self-control (cf. Christman: 2015; Oshana: 2006). Without an understanding of one's motives, desires, and actual beliefs, it will not be possible to ask whether one really wants to be moved by them: hence, there is a need for self-reflection. The whole idea of being autonomous involves more than just governing one's life: to some minimal extent, one needs to govern one's life *well*. This is why the autonomous person engages in the following three modes of rational thought (Baumann & Döring: 2011). She finds appropriate means to reach her ends. She will also seek to maintain a sufficient degree of coherence within the entire complex of her aims, values, commitments and circumstances in life. Otherwise, her efforts to direct her own life will be in vain because of either conflicting desires or aims that are too difficult to reach given external constraints. Most fundamentally, she will consider which ends are worth pursuing in the first place. Finally, the self-reflective and rational person who knows how she wants to live still must follow up on that judgment by actually living that life: she needs a capacity for self-control.

Having some minimal capacity for autonomy is generally thought to confer the '*right* to autonomy' on people in society. This right to autonomy is particularly relevant for our discussion of the interaction between PTs and users. It includes the right to be free from controlling interference by others, such as manipulation, deception, and coercion. Each of these interferences threatens or obstructs a person's ability to govern her own life. Typically, all adults have this right to autonomy unless they have some severe condition that justifies restrictions. There is widespread agreement in Western civilized societies and, more specifically, in applied ethics that personal autonomy is a central and important value that demands protection and respect.[11] Nevertheless, it is also commonly acknowledged that there can be overriding (moral) reasons that justify interfering with autonomy. Given the importance of autonomy, these reasons must be

[11] Different ethical theories vary in their reasons why autonomy is such an important value. Kantian theories emphasize autonomy as a characteristic of the will that distinguish us as humans. Respect for autonomy thus means respect for persons; autonomy is regarded as intrinsically valuable (Kant: 2003). Consequentialist theories often regard autonomy as a means to other goods. For example, Mill thinks that people themselves know best what will make them happy. Hence, granting people a strong right to autonomy is a very good means for achieving the greatest happiness (Mill: 1985).

both substantial and compelling. Hence, the present discussion grants users of PT a relatively strong right to autonomy, leading to clear constraints on what designers and deployers are allowed to do without user informed consent.

PT scholars emphasize that PT should leave users free to decide and act for themselves because they recognize PT's immanent threat to autonomy. As noted above, this threat can take several forms, three of which I now will discuss and address. First, as explained by Fogg (2003, 7–11, 213–218), the power balance between PT and users can become asymmetric to the extent that it threatens users' freedom to behave as they believe they should. The fact that PTs can be easily designed to control communication with users is a major source of power asymmetry. The manner in which PTs are programmed by their designers determines and limits human users' possibilities for interaction. In human-human persuasion, each party can stop the interaction, can always ask for clarification, and can show in numerous ways that she feels uneasy with the persuasion process. In contrast, currently existing PTs often remain limited in their capacities for this sort of two-directional communicative interaction. On a related note, PTs can also be "proactively persistent" (Fogg: 2003, 216) because unlike humans, they do not become tired or embarrassed, nor start feeling uneasy or guilty. PTs can continue their persuasive attempt until the user capitulates in either a moment of weakness or a moment of unawareness. Another gap in user control arises from the fact that PTs are not yet sufficiently able to reliably detect and respect user emotions (cf. Baumann & Döring: 2011), whereas they can show (programmed) emotional expressions, which can be a powerful means of persuading humans.

PTs can have a wider range of access, which is a further source of power asymmetry. Many PTs can go where humans are not allowed to go—e.g., in the bathroom or in the car—thus enabling them to persuade at the right moment and place. Also, by means of its impersonal appearance, a PT may suggest anonymity and bypass the social barrier to certain topics in human-human persuasion. Finally, power asymmetry may arise from the fact that PTs (being or embedding computers) generally have great capacities. Computers can "store, access, and manipulate huge volumes of data" (Fogg: 2003, 8), which enables them to provide the right piece of information at the right place and time to persuade. Once successful persuasive strategies are developed, their application potential can be multiplied with the help of PTs. Based on their substantial processing capacities, computers can also employ several modalities—to wit, audio, video, text, graphics, animations, and hyperlinks can be used and combined to enhance and tailor persuasion.

The threat to user autonomy posed by power asymmetry should be addressed in the following ways; although they are primarily the responsibility of designers and deployers, there is also a role for users and society. I suggest that as a prima facie guideline, users must be free to decide whether to be the subject of persuasive communication by a PT, and if they choose to use the PT, they must be given the option to switch it off whenever they wish to do so. In this way, users can control the times at which they interact with PT and for which of their aims, in which spaces, and to which means of technological persuasion they are willing to be subjected, thus safeguarding their autonomy. The option to mark certain online advertising as unwanted, for example, fits this guideline. In some cases, especially when a user needs a PT to achieve self-chosen aims, he might prefer to continue interacting with the PT even if he feels pressured by the PT. Therefore, it would be too binary to provide nothing but the option to switch off the PT. Instead, designers must give users control over the relevant settings of the PT. In case of health apps, for example, users should be given the option to set time windows in which they do not wish to receive reminders and suggestions, the option to determine how often they receive feedback, the choice of whether to participate in a social community of app users, and so on (for further discussion, cf. (Nickel & Spahn: 2012)).

One example of a consideration strong enough to override the prima facie guideline might be patient safety in a hospital in the case of the *Hygiene Guard*, which warns medical personnel of the dangers of failing to properly wash their hands (Fogg: 2003, 47). Alternatively, perhaps we democratically decide that sustainability is important enough that it justifies imposing a duty upon each driver to continuously run eco-feedback technology (cf. Somsen: 2011). In such cases, autonomy is outweighed by another value—in the first example, safety, and in the second example, sustainability.

So far for the discussion of power asymmetry. A more narrowly circumscribed threat to user autonomy concerns, secondly, a PT's so-called means of persuasion. In part, a PT will employ factual information to persuade users on the basis of information and arguments. In addition, however, PTs use what I will call 'non-argumentative means of persuasion' (see Chapter 3). Here, one can think of emotions or several forms of social influence, such as peer pressure, appeals to authority, and the like. Since we often lack awareness of the operation of the cognitive mechanism by which these non-argumentative means of persuasion exert their influence on us, there is a clear danger of manipulation.[12] In other words, it might be that non-argumentative means of persuasion subvert or supplant our deliberation and decision-making. To address this threat to autonomy, designers and deployers must carefully ensure that their means of persuasion grants users 'substantial control' over their mental states and behaviour. One way to achieve this is by informing the user about the persuasive methods used ('transparency of methods', (Compen, Spahn, & Ham: 2015, 225)). In the next two chapters, I will more thoroughly investigate this topic of a PT's means of persuasion in relation to a user's substantial control.

At this stage of our discussion, however, it is worth considering the phenomenon of 'personalized persuasion'. This phenomenon involves tailoring the chosen means of persuasion to the individual user, thus increasing the chances of persuasive success (Berkovsky, Freyne, & Oinas-Kukkonen: 2012; Fogg: 2003). For example, consider personalized recommendations made by web shops, such as 'your friends like this X'. In the course of interacting with a user, a PT can determine the user's 'persuasion profile', i.e., his relative susceptibility to different types of influence (Kaptein & Eckles: 2010, 86). By using its huge data processing capacity, the PT can do this for all its users. This strategy will be particularly successful when different deployers of PTs share such information (hence the need to consider privacy, as set forth below). Designers and deployers have a duty to ensure that persuasive profiling does not deprive users of substantial control over what they decide. This includes considering the possibility that some users are unusually perceptive to some type of influence. Another duty is to ensure that personalized recommendations or behavioural suggestions are indeed personalized. Thus, if advice purports to be tailor-made for an individual user, it must actually be tailor-made; otherwise, it is deceptive and manipulative.

The more successful persuasive profiling is, the more reason to worry about user autonomy; thus, there might be inherent moral limits to the pursuit of this strategy. It might very well be the case that using personalized PT requires deployers to ethically adapt PT to each individual user. Personalizing PT might also entail a duty to personalize the application of ethical principles, such as the non-manipulation principle. For example, it would be conceivable that a com-

[12] I do not hold that the only way humans can control what they think and do is by conscious reflection. Rather, I hold that regularly we do not control how non-argumentative means of persuasion influence us, *partly because* they operate under the radar. For further discussion, see Chapter 3 below.

pany could gather sufficient data on an individual consumer's interaction and purchase history to conclude that this consumer nearly always follows up on a recommendation based on what his friends bought. This seems to be so clearly irrational that prima facie, the most plausible explanation would be an irrational fear of 'missing out' or exclusion from one's peer group. Playing on this irrational type of motivation seems to represent a clear case of manipulation (cf. Cave: 2007). Nevertheless, it is morally permissible to give product recommendations to the average consumers based on his peers' purchases.

A third threat to user autonomy involves privacy.[13] Although privacy is a multifaceted concept, in the context of PT, the most relevant type of privacy is defined as 'control over information about oneself' (DeCew: 2015). If other parties illegitimately gather sensitive information about me, knowing that to do so might make me feel less secure about governing my own life, my autonomy will be jeopardized. If, unbeknownst to me, those parties use that information to influence me, they are manipulating me, and I am not governing my own life. A great deal of privacy-sensitive data is generated during PT-user interaction. Examples include information about a person's health, interests, whereabouts, financial position, political preference, consumer preferences, and so on. Since this information is marketable, deployers of PT have an interest in this information (cf. Huckvale, Prieto, Tilney, Benghozi, & Car: 2015). Here, the general guideline is that deployers are not allowed to gather, store, and sell privacy-sensitive information without informed user consent. Users can be considered to give informed consent when the following conditions are met: they are given the relevant information, possess the relevant capacities to make decisions, and give permission voluntarily (Eyal: 2012; Faden & Beauchamp: 1986). Limiting our attention to the first condition, this means that deployers must make clear what sort of information they wish to gather, the purposes for which they intend to gather it, and the parties with which they will share it.

So far, this discussion has addressed several threats to user autonomy. However, PTs are often designed with the aim of enhancing autonomy. In other words, they are designed as self-help tools for users to change their behaviour in important ways, e.g., to quit smoking (Dijkstra: 2006; see also 4.3.5 below). Such PTs are 'scaffolding' technologies, helping users overcome problems of limited

[13] While there is debate on the question whether privacy is foremost connected to autonomy, or to some other value, most authors hold that privacy is an important good, which we should care about (DeCew: 2015).

willpower and self-control (Anderson & Kamphorst: 2014). In the case of this class of PT, some of the guidelines given may be applied differently. So, for example, it would be less problematic if certain means of persuasion operate ‘below the radar’, provided the decision to employ them is authorized by users. Hence, as in the case of privacy, informed consent is crucial. Designers and deployers should note, however, that tacit consent to manipulative influence strategies cannot always be inferred from a user’s decision to accept help from a PT. Therefore, designers and deployers must explain their methods and ask users for explicit consent.

In this section, I have briefly discussed how to address various ways in which PT endanger personal autonomy. In the next section, we will turn to user behaviour and its typical consequences for several values other than autonomy.

2.3.2. User behaviour, target value, and other values

Above, the distinction was made between user behaviour and its consequences analysed in terms of how various values are affected. This will prove fruitful in analysing several ethical issues that pertain to user behaviour and the PT’s end. First and obviously, ethical attention is required with regards to a PT’s goal or end (or, what I call ‘target value’) (Berdichevsky & Neuenschwander: 1999; Fogg: 2003; Schermer: 2007). In terms of Berdichevsky and Neuenschwander, along with Fogg, the “intended outcome” of the PT should not be unethical, and the PT’s goal should be transparent to the user. For example, a PT designed to enable parents to “persuade children to divulge their secrets” would be unethical (Berdichevsky & Neuenschwander: 1999, 53). One could also envision the excessive use of persuasive strategies in games to ensure that players continue to play, potentially leading to compulsive gaming. This is even more problematic if it is unclear to gamers that the game is intentionally designed to achieve just that.

In terms of the diagram, designers should ensure that the PT’s end could be appropriately referred to as ‘target *value*’ and thus is a valuable end indeed. To prevent misunderstanding, this does not mean that a PT is required to have an idealistic or morally supererogatory end; instead, it means that it is not morally wrong to pursue that end and if that end is realized, some value is realized. Thus, PT designed to sell consumer goods has the target value of ‘economical profit’, which is morally acceptable in itself. Given the importance of user autonomy, and given the idea that a PT relies on voluntary user participation, users should be able to endorse (or at least to accept) the PT's goal. In addition, it is also for the user to determine how important the PT's goal is relative to the other things they value in life.

In some cases, the nature of the PT's end might justify allowing designers and deployers to work with users in ways that typically are not permitted. One case is briefly noted above: sometimes PTs are designed to prevent harm to others. Since harm to others can be a ground for limiting user liberty (Feinberg: 1987), mandatory use of such a PT might be justified (Smids: 2016). In addition, Berdichevsky and Neuenschwander claim that designers have no duty to disclose the intended outcomes (i.e., the PT's end) "when such disclosure would significantly undermine an otherwise ethical goal" (1999, 53). However, this is too simple, as first, we need to know more about that 'ethical goal'. If, for example, the goal is for smokers to quit, then this goal, although valuable, regards the good of the user himself. The issue of whether the user truly wants to quit smoking is within the bounds of his personal right to autonomy. Hiding the PT's end would be paternalistic, substituting the user's judgment of what is best for him, all things considered, with the designer and deployer's judgments. Generally, given a user's status as an autonomous citizen, such paternalism is unjustified, whereas limiting liberty to prevent harm to people other than the user can often be justified (Dworkin: 2016; Feinberg: 1987, 1989).

With regard to the PT's target behaviour, which should help achieve the PT's target values (its end), there are also ethical issues. First, it is often crucial both that the PT successfully leads users to engage in the target behaviour *and* that this sufficiently contributes to achieving the PT's end. Taken together, this means that the PT should be effective, given a user's voluntary participation. Suppose someone chooses to use a weight-loss website or app (such as 'Calorie Count'[14]) to lose weight and live healthier. He invests money, time, and hope in using it. Therefore, he wants the site or app to successfully persuade him to act as he is required to act. Furthermore, he has a legitimate expectation that following up on the site or app's advice and behavioural suggestions will lead to losing weight and living healthier.[15]

[14] https://www.caloriecount.com/

[15] His expectations are legitimate because the site or app claims to help users to lose weight if they use it in the right way. In case the user also pays for the service, his expectations are all the more justified.

A PT's effectiveness and quality are by no means guaranteed. The first hurdle, persuading the user to engage the target behaviour(s), is not always (fully) cleared (Hamari, Koivisto, & Pakkanen: 2014). Most likely, the major challenge is the configuring of the exact set of target behaviours for each individual that lead to the desired end. To follow up on the previous example, studies investigating several weight-loss apps found that they were not sufficiently grounded on weight-loss practices that are informed by scientific evidence (Azar et al.: 2013; Breton, Fuemmeler, & Abroms: 2011). This is a clear reason for worry. In addition, there is an important issue regarding the accuracy of data input and corresponding advice (Voerman: 2015). Although users must report every instance of food and drink consumption, even if they do, there will always be a substantial margin of error, with the result that the user's total calorie intake is never more than an estimate. Similarly, the app's calculations of daily calorie use depend on very rough descriptions of physical activities (e.g. '30 minutes walking with moderate intensity'). Consequently, it is quite possible that the app counts a net calorie intake lower than the numbers burned, incorrectly informing the user that he is doing well. Therefore, it cannot be taken for granted that the performance of the target behaviour furthers the PT's target value.

Ineffective PT is a serious problem since individual users, other stakeholders and society at large make a substantial investment in using PTs. If a PT is the government's choice to address social problems such as obesity, designers and deployers should be able to tell whether the PT has the potential to be cost-effective.

Even if a PT is effective in the sense that the target behaviour fosters the target value, this may (as visualized in the diagram) not tell the whole story. The weight-loss example is an apt illustration (cf. Schermer: 2007; Verbeek: 2006). Suppose calorie counting (target behaviour) leads to weight loss, having a positive effect on health (target value). However, counting calories may also make eating more complicated and stressful, thereby reducing the sociability of eating with others. This is an example of the target behaviour having a negative impact on other values (of course, the social function of meals also contributes to health). Of course, the target behaviour might also have a *positive* impact on other values. For apps such as a calorie counter, these values might not be immediately identifiable. An example from another context is eco-feedback technology in cars that supports smooth driving, resulting in reduced fuel consumption (Barkenbus: 2010). In addition, eco-driving behaviour improves traffic safety by reducing both the incidence of high speeds and speed differences between cars (Barth & Boriboonsomsin: 2009).[16] However, the target behaviour also includes the driver's monitoring and interpreting eco-feedback. If this behaviour causes driver distraction (cf. Verwey, Brookhuis, & Janssen: 1996), the safety gain will be smaller.

One example of unintended behaviour is the following. Weight-loss websites or apps may have the unwelcome result that users over-interpret their health in terms of the behavioural suggestions, undervaluing other constituents of good health. For example, users might not pay sufficient attention to their bedtimes and fail to get enough sleep. In that case, part of the gain in health by losing weight is undone; the target behaviour and the unintended behaviour have opposite effect on the target value.

These examples clarify that it is a complex matter to assess a PT's real-life impact. First, one has to determine, or to predict as best as possible, the target behaviour and various (possible) unintended user behaviours. Next, one has to search for positive and negative effects on both the target values and several other values. The examples discussed above should suffice to show that making these distinctions extends our analytical toolkit compared to merely distinguishing between intended and unintended outcomes.

The question left is how to deal with cases in which a PT negatively affects some other values. Note that asking this question presupposes both that these negative effects are clearly in focus *and* that something can still be done. However, it is relatively unlikely that both are true at the same time because negative effects are most visible in the stage where the least can be done, that is, when the PT is fully in use in the real world (Collingridge: 1981). Nevertheless, suppose that deployers and designers more or less accurately forecast a value conflict in the sense that realizing the target value comes at some cost to other values. Ideally, they manage to mitigate negative effects by altering the design of the PT. In cases where this is not possible, it must be decided whether positive effects on the PT's target value outweigh negative effects on other values. It is crucial that the various agents involved in design, deployment and use participate in the deliberation. Given the user's right to autonomy (section 2.3.1 above), if the PT's negative effects concern them in the first place, the final say in making such value trade-offs belongs to them. Because such value conflicts are the rule rather

[16] And most likely, it will also reduce traffic congestion.

than the exception, there is good reason for always involving users in the design and development process.

In conclusion, we also need long-term studies of how PT fosters or negatively affects the values at stake in PT use. The impact of a PT may change over time. Unfortunately, few long-term studies have been done (but see e.g. Kampf: 2016) since much research is done in experimental settings that involve limited prototype testing.

2.3.3. Deployers and designers

Deployers and designers together largely determine what PT will be developed and thus the extent to which it will be ethically sound and will contribute to the good of both users and society. However, their differences in knowledge, power, and interests may give rise to problems. Imagine a company (the deployer) that asks a team of designers to build a very effective web-shop, one that is highly successful in persuading visitors to buy products. If the designers simply accept the commission, they might run the risk of assuming all responsibility for creating a PT that is *both* successful *and* ethically sound, for example, one that respects user autonomy and privacy. However, the tension between these objectives and the pressure to deliver might tempt the designers to compromise on user autonomy. This worry is even more significant because the distinction between persuasion and manipulation is not always easily drawn (see section 4.4 below), which could make it appear attractive or less problematic to enter into morally grey areas.

The best way to prevent problems such as those set forth above is for deployers and designers to assume co-responsibility. Deployers and designers each have moral and professional duties to ensure ethically sound PT. First, and most fundamentally, as humans, they ought to treat their fellow humans in ways universally recognized as moral, e.g., having sufficient regard for their well-being, not deceiving them, etc. In addition, deployers have legal duties, such as those arising out of the legal regime of product liability. Designers will often be bound to professional codes of conduct, for example, the US National Society for Professional Engineers' 'Code of Ethics for Engineers'. This code urges engineers to "[h]old paramount the safety, health, and welfare of the public" (National Society of Professional Engineers: 2007). Because deployers and designers have different expertise and roles in developing a PT together, they depend on each other to be faithful to these duties. Therefore, they have no choice but to cooperate and together take responsibility for morally sound PT. I will discuss that cooperation in the context of a few ethically sensitive issues.

First, to follow up on the web-shop example, deployers need to mandate and even require designers to incorporate only means of persuasion that leave users with substantial control over their behaviour. They need to operate from an explicit, shared understanding of the importance of user autonomy. Deployers are dependent on designers, who have the expertise necessary to determine how the means of persuasion affect user freedom. For their part, deployers may have more knowledge of the intended users and the use context, which may be relevant for evaluating the PT's impact on user autonomy. In cases in which informed user consent justifies further-reaching means of persuasion, designers and deployers must cooperate both to provide adequate information to users and to obtain consent.

Second, cooperation is indispensable with regards to the quality and effectiveness of the persuasive technology, along with its adaptation to individual users. As shown by the example of weight-loss apps, the quality of a PT cannot be taken for granted. Consider the class of self-help PTs, which users actively choose as a supporting technology for reaching goals, such as losing weight. Users will rely on the PT's information, advice, and behavioural suggestions. The options suggested by a PT will significantly influence which options a user considers (Kamphorst & Kalis: 2014). As noted above, the user will often assume that these options are valid, are based on the best available knowledge and if followed up, will be effective in helping them achieve their goals. Indeed, it is conceivable that users will interact extensively with a PT, may rely on a PT and sometimes might even be grateful for a PT's positive impact on their lives. For that reason, it makes sense to describe users as trusting the PT they use (Nickel, Franssen, & Kroes: 2010; Nickel & Spahn: 2012). Given that this class of self-help PT is generally promoted as an aid to achieve self-chosen aims, we should, prima facie, regard user expectations and trust as legitimate. Living up to these expectations can impose a relatively heavy burden and again, designers and deployers need to cooperate by sharing expertise and enabling each other to fulfil his specific role.

One complication is that to a substantial extent, the quality and effectiveness of PT are related to individual users, who may differ significantly in the dimensions of their physical make up, personal history, etc. Especially in the medical domain, health apps and e-coaches need to be designed to adapt to individual users to avoid misfiring and becoming detrimental to health. The suggestion to go for a 5-km walk may be appropriate for one person, but far too ambitious for others. A specific domain such as health generally increases the responsibility of designers and deployers in their professional roles. They must be aware that their PT will significantly shape the user's normative beliefs of what is healthy, wise, normal, etc., and will have a significant impact.

Third, deployers and designers need to cooperate to predict, prevent and or mitigate a PT's cost to other values. Here again, deployers may have a better or additional insight into who the users are and the context in which they will use the PT. By sharing that knowledge with designers, by involving users in the design phase, and by investing in serious prototype testing, it is the most likely that a PT's potential negative effects will be resolved.

To conclude, the three issues briefly discussed above illustrate the value of focusing on designers and deployers together instead of designers alone. Thus, adding 'deployers' to Berdichevsky and Neuenschwander's diagram makes a fruitful contribution to the ethics of persuasive technology.

2.3.4. Ethical issues that emerge in the wider context

There are a few important ethical issues that can best be discussed by taking all agents and the wider context into account. This is because it is sometimes less immediately clear how an issue arises from a complex interplay between many actors and causes and who is primarily responsible for addressing that issue. I will now discuss three such issues, considering the roles of the user's significant others (relatives, friends, etc.), the government, regulators, interest groups, and the like.[17]

First, the availability and use of PTs raise a question regarding justice. Many PTs are Internet based, or are smartphone apps, or involve other ICT. Unfortunately, the tech-savvy are better positioned than other people (especially the elderly) to reap the benefits of using PTs. Governments make a substantial investment in developing PTs such as e-coaching to combat several social ills, such as obesity, depression, etc. Even if such strategies are hugely successful in the aggregate, governments must insure that nobody is excluded. This could be done by demanding and requiring deployers and designers to provide additional

[17] Note that taking the wider context into account is relevant and often helpful for all of the ethical issues treated in this chapter. For example, the context might specify or modify designer's responsibility, explain why some users are particularly vulnerable, etc.

user support to citizens who lack critical capacities or attitudes. In addition, governments should ensure that non-PT based support remains available for such citizens.

Second, there is a concern that the extensive and widespread use of PT means delegating too much of what makes us human to technology (Anderson & Kamphorst: 2014; Spahn: 2015; Verbeek: 2009). We are witnessing an historical transition from living without ICT applications to living in an environment that is pervaded with ICT applications ranging from large to small, from visible to hidden (Aarts & Wichert: 2009; Nordmann: 2004). It is to be expected that PTs will become increasingly integrated into that environment, trying to persuade us to engage in various patterns of behaviour that would be good for us (Kaptein et al.: 2010). However, when a PT suggests how to behave in a specific circumstance to achieve a goal, merely by following up, we can avoid relying on our own judgment to determine the best course of action. Thus, we do not engage our own capacities for moral judgment and decision-making. In the end, we may delegate too much of our moral judgment and decision-making to PTs and weaken our capacities to an extent that threatens our distinct humanity.

The same concern relates to our capacity for self-regulation. Even if we make up our mind about what is good for us to do now, we do not always manage to follow up. We sometimes lack willpower altogether, or our resources for exercising will-power have been depleted (Muraven & Baumeister: 2000). Often PTs are designed to support us in overcoming this weakness of will, for example, by giving users peer support via online forums. However, part of what makes us human is the ability to strengthen our capacities for self-control by exercising them. In this way, we develop good habits and may acquire virtues. Clearly, relying too much on PT to circumvent problems of self-regulation undermines these efforts. Closely related is the importance of having a sense of ourselves as authentic persons. It is important that we experience ourselves as the authors of our actions. If we feel that it is actually PT—not us— that causes us to eat more healthily, our self-esteem and sense of self-efficacy will be diminished. This would have the adverse effect of decreasing our general ability to successfully pursue our goals (Ajzen & Fishbein: 2000; Anderson & Kamphorst: 2014; Anderson & Kamphorst: 2015).

These ethical issues are far-ranging, multifaceted, complex, and contested. Therefore, they require extensive public debate, which ultimately ivolves the fundamental question of what it means to live a good life (cf. Verbeek: 2009). Governments should foremost initiate and facilitate this debate both because of its fundamental nature and because citizens' long-term well-being is at stake.[18] In the meantime, some recommendations can be made to develop and use PTs in such a way that we remain human. Under-exercising various of our human capacities is the theme that unifies the concerns discussed above. Naturally, part of the solution should be sought in designing PTs that support our capacities for judgment, decision-making, and self-control in ways that involve and engage these capacities, rather than displace them.

Following are examples of some design suggestions for engaging the user's relevant capacities. PTs could involve the user in goal-setting, instead of one-sidedly determining, based on various considerations, what (sub)goals the user should achieve to reach his overall aims (such as a healthier life, sustainable driving, etc.). In this way, users must deliberate and can calibrate their PT-related goals with their other goals and projects. Another design suggestion is to give users choice in the type and frequency of suggestions, feedback, motivational support, etc. In that way, users can choose to receive only the level of support they need.[19] Ideally, the PT regularly asks users to reconsider their settings. Alternatively, after registering some level of successful goal-achievement, the PT could suggest scaling down its behavioural support, thus stimulating users to strengthen their own capacities. Still another thing developers might do is to phrase PTs' feedback and behavioural suggestions in ways that stimulate user reflection and decision-making. For example, suppose it is 9 PM and the user of a weight loss app has not yet done her target amount of physical exercise. Instead of giving the highly specific suggestion to go for a 20-minute walk right now and telling the user that she will feel proud, happy, etc., the app could merely ask her whether she knows she has not exercised, leaving it to her to deliberate and decide whether, when, and how to do so.

A third emergent ethical issue involves the problem of assigning responsibility for ethically sound PT to the various parties involved in its development and use (Schermer: 2007; Verbeek: 2006). Looking again at a persuasive health app, these parties include designers, software developers, medical advisors, manufacturers, medical caregivers, regulators, technical support staff, users, and

[18] In the Netherlands, this task is firmly taken up by the publicly funded Rathenau Institute, resulting in debates, workshops, lectures, and publications such as (Kool, Timmer, & Est: 2015b).

[19] See (Kaipainen, Mattila, Kinnunen, & Korhonen: 2010) for a nice example of a PT that exhibits several of this recommendations.

relatives, together forming a complex network. If something goes wrong, the worst-case scenario would result in death or severe disease. Often, it will be difficult to determine each party's causal contribution to the tragic event. In addition, it will be difficult to determine the extent to which each party possessed the knowledge and expertise by which that party could have prevented it. This knowledge problem is particularly pressing in the case of PT, crucially turning on human behaviour, which is notoriously difficult to understand and predict. In addition, the PT may be used either outside of its intended context or by unintended users (Fogg: 2003, 229). In such conditions, the so-called the 'problem of many hands' arises, which has been defined as "a gap in a responsibility distribution in a collective setting that is morally problematic" (van de Poel, Fahlquist, Doorn, Zwart, & Royakkers: 2011).

There are a few reasons that it is morally problematic for each actor's exact moral responsibility to be unclear. If something goes seriously wrong, users or their relatives will reasonably want to know which party or parties should be held accountable and if appropriate, blamed for their shortcomings or wrongdoings. In addition, it should be clear when legal punishment is legitimate and if (financial) compensation is owed to the victims, who should pay. Whereas these reasons all relate to so-called 'backward-looking responsibility', the problem of many hands also applies to 'forward-looking responsibility (van de Poel: 2011; van de Poel et al.: 2011). Forward-looking responsibility means that each party involved in the development and use of a PT has certain role-specific obligations. In other words, each party has obligations to ensure that the resulting PT and its use in daily life conform to legitimate expectations of quality, reliability and the prevention of negative outcomes. The distribution and assignment of these role-responsibilities must be adequate to deliver ethically sound PT.

Fortunately, some measures can be taken that together should both bring more clarity about each agent's role responsibility and expand it where necessary. Initially, governments and regulating bodies should set quality standards for PTs (Kool, Timmer, & Est: 2015a). Fulfilment of these standards might be communicated to users by means of quality certificates. This is particularly important for PTs that do not yet qualify as medical technology and thus are not regulated by the practice of medicine but nevertheless aim to improve user health and well-being. For such a PT, the quality certificate might include the demand that the PT's information, feedback and behavioural suggestions be evidence based. An additional measure is adequate prototype testing, which should be part of the set of quality standards. Such prototype testing is indispensable to investigate how users use the PT and may reveal either unintended behaviour or instances of costs to other values caused by the target behaviour. Another way to gain the necessary input from users is participatory design, i.e., the involvement of users in the design phase (Davis: 2010).

The final measure regards the communication and transfer of responsibilities between scientific researchers and (teams of) designers. As can be inferred from the proceedings of the yearly conference 'Persuasive', a substantial amount of scientific research on PT is quite practice oriented, yielding tangible results ranging from a 'proof of principle' to relatively developed prototype PTs (cf. Bang & Ragnemalm: 2012; IJsselsteijn, de Kort, Midden, Eggen, & van den Hoven: 2006; Ploug, Hasle, & Oinas-Kukkonen: 2010; Spagnolli, Chittaro, & Gamberini: 2014). The concern is that scientists, often already engaged in prototype testing, pay insufficient attention to ethical requirements simply because their primary aim is not to develop a ready-to-market PT that is ethically sound but instead, to extend our scientific knowledge of PTs. Designers and product developers might simply further develop the prototypes without sufficient attention to ethical issues. Moreover, they probably will have less intimate knowledge of the PT's means of persuasion and its interaction with users than do scientists; to a certain extent, they may treat the prototypes, or parts of it, as black box technologies. One possible way to prevent these problems is to involve scientists in the ethical assessment later in the design phase. Another is to require or encourage scientists to pay more attention to ethical questions, especially the issue of how their means of persuasion relate to user autonomy.

On a concluding note, if we reflect on the issues discussed in this section, we can make an important observation: one issue *is* the problem of many hands, whereas the others—justice and preserving our distinct human capacities—are more difficult to address *because of* the problem of many hands. This means that we cannot leave it to the market or other private parties to address these issues (even though we identified several clear, rather far-reaching responsibilities of designers, deployers, and scientists). These issues transcend individual parties' immediate interests and responsibilities. Given the urgency of the issues and the public interest, the government should assume a role as the regulator and facilitator of the requisite cooperation among the many agents involved in the development and use of PTs.

2.3.5. Summarizing overview

Numerous ethical issues have been discussed in this chapter. Table 2.1 highlights several of them. When read horizontally, for each agent in the left column, e.g., ‘deployer’, the table specifies his moral obligations with regard to the designer, deployer, PT, user, etc. The table can also be read vertically, specifying for each of the subjects designer, deployer, PT, etc., the responsibilities of other agents towards this subject.

Responsibility of:				
Government and regulators (part of context)	**User**	**Designer**	**Deployer**	**With regard to:**
Specify responsibilities	Give or withhold informed consent to: methods of persuasion use of privacy sensitive data	Cooperate with deployer; assume co-responsibility for ethically sound PT Maintain professional independence	Adhere to common and personal morality, the law, and corporate code of conduct	**Deployer**
Specify responsibilities	Cooperate with prototype testing and provide consumer feedback	Adhere to common & personal morality and professional code of conduct	Cooperate with designer; assume co-responsibility for ethically sound PT Enable designer Involve users in design process	**Designer**
Set standards for effectivity quality, privacy, informed consent. Ensure efficacious supervision	Take effort to shape the interaction with the PT	Incorporate values in design: autonomy, privacy, safety, functionality	Monitor user-PT interaction	**PT**
	Protect personal autonomy and human dignity	Grant users sufficient control over PT, their mental states, and behaviour	Respect user autonomy, privacy, and other values Obtain informed consent when required	**User**
Monitor occurrence of UB and negative effects on TV and OV. Special attention to 'emergent issues' (justice, responsibility, dignity)	Reflect on performing TB Beware of UB and its effect on TV and OV.	Ensure that TB fosters TV (given user cooperation) Predict UB and negative effects on TV and OV.	Anticipate and monitor occurrence of UB and negative effects on TV and OV.	**Target behaviour (TB)** **Unintended behaviour (UB)** **Target value (TV)** **Other values (OV)**
		Show sensitivity to: Context of daily use (user, friends, private vs public, etc.) Interests of society at large	Abide by the relevant laws Show sensitivity to: Context of daily use Interests of society at large	**Context:** **-government** **-regulators** **-relatives & friends of user**

Table 2.1 Overview of points of ethical interest, ordered by agent in relation to the various nodes of diagram 2.3

2.4. Concluding remarks

In this chapter, Berdichevsky and Neuenschwander's diagram was modified and extended in several ways to facilitate ethical reflection on the development and use of persuasive technologies. The most important modification and extension concerns the ambiguous concept of 'outcomes', which was substituted for with the distinction between behaviour and the effects of that behaviour. These effects can be evaluated in terms of the target value, that is, the end for which the PT was designed, and other values. By adding the distinction between target behaviour and unintended behaviour, the new diagram enables a more fine-grained analysis of a PT's impact. The various agents involved with PT become better equipped to trace the origin of undesirable effects of PT and thus to address these effects.

Another important addition to the diagram, viz. deployers, was one step to remedy the gross simplification that it is only the designer who develops and implements a PT. I argued that designers and deployers should assume co-responsibility for ensuring ethically sound PT. Adding deployers also helps to better connect designers to the use practice of PTs since it is the deployer who actually implements the PT in daily use.

A major finding from this chapter's overview of the ethics of PT is the observation that governments play a crucial role. First of all, they must ensure that the responsibilities of individual parties are clear and sufficient. But they must also ensure that emergent ethical issues, such as justice and the preservation of our distinctive human capacities (for moral judgement, decision-making, and self-control), are adequately addressed. This will include generating public debate on these fundamental topics and facilitating designers, deployers, and other actors to address them.

Each of the ethical issues identified in this chapter deserves much further study. A few will be the subject of later chapters of this thesis. In the next two chapters, I will elucidate the question of how we can determine whether a persuasive influence strategy gives the user substantial control over her mental states and behaviour. Whereas this chapter was relatively general, ethical reflection on PTs will benefit from case studies. Chapter 6 contains a detailed study of the persuasive version of Intelligent Speed Adaptation.

Finally, the question arises whether this chapter's approach to the ethics of persuasive technology suffices for safeguarding ethically sound PT. It has become evident that PT is an ethically sensitive technology because of its very nature: PT aims to influence human behaviour. Consequently, there are many ways in which a design must adequately consider the moral values at stake. However, looking at the proceedings of the yearly conference on PT, we seem to encounter the following situation (see references in 2.3.4 above). A few ethicists seek attention for ethical issues; although most psychologists and designers sincerely acknowledge these issues, in general it seems as though they do not account for those issues in PTs' research and design.[20]

Part of a solution would be for scientists (developing prototypes) and designers to seek help with design approaches that include ethics in the design process. Previously, I mentioned the Participatory Design method. Another such approach is Value Sensitive Design, which consists of a three-stage methodology that aims at designing in such way that crucial ethical values are realized as much as possible in design (Friedman, Borning, & Huldtgren: 2013; Friedman & Kahn: 2003; van de Poel: 2009). If ethically sound PT is to be realized, it is the scientists, designers, and developers who *invent* the PT who play the crucial role (not ethicists who merely *write* about the ethics of PT). It is part of their role responsibility to take up this challenge.

[20] For one of the exceptions, see (Reitberger, Güldenpfennig, & Fitzpatrick: 2012)

3 Should we prefer nudging over rational persuasion?

3.1. Introduction

Governments have powerful reasons to attempt to influence their citizens, e.g., to promote health and safety. But how? In the last decade, 'nudging' has been advocated as a new and effective approach to influencing people (Sunstein & Thaler: 2003; Thaler & Sunstein: 2009).[21] Nudging refers to deliberately re-arranging a decision-making situation in such a way that desirable choices become more likely. For example, consider placing healthy food at eye-height, increasing the visibility of stairs over that of the elevator, and the like. More specifically, nudging is the use of psychological insights into the characteristics and imperfections in human deliberation and decision-making to "alter people's behaviour in a predictable way without forbidding any options" in a way that should make them better off (Sunstein & Thaler: 2003; Thaler & Sunstein: 2009).

Proponents of nudging and critics who favour rational persuasion differ over the question how governments should influence citizens. Nudge proponents argue that setting up the decision-making context in some way (choice architecture) is inevitable and therefore should aim at making citizens better off. The critics of nudging argue that it often amounts to manipulation (Grüne-Yanoff: 2012; Wilkinson: 2013) and that only rational persuasion fully respects citizen autonomy (e.g, Hausman & Welch: 2010).

This debate suffers from several shortcomings that the present chapter aims to address. Nudge proponents base their case entirely on the findings of social psychology and behavioural economics by arguing that nudging is the best response to these findings. Proponents of rational persuasion, however, do not engage with these findings. They define rational persuasion in a way that pres-

[21] The term 'nudging' is most well-known, but the same approach is, advocated by other authors as well, albeit under different names. e.g. (Trout: 2005). As explained in in the Introduction, nudges that involve technology as a means for influencing have similarities with persuasive technology.

umes that recipients are capable of changing attitudes, preferences, and behaviour based *solely* on unbiased reflection on arguments.[22] The problem, of course, is that recipient reflection often will be shaped by the same heuristics and biases that occupy centre stage in behavioural economics (see, e.g., Kahneman: 2003).

In this chapter, I will argue that a rational persuasion-based approach to influencing citizens remains viable, also in the light of the findings of the behavioural sciences. This is because that approach can indeed account for and even make *use* of the findings of behavioural economics and psychological persuasion research. Therefore, nudging is not at all inevitable given our increased knowledge of human psychology. In addition to establishing the viability of rational persuasion, my further aim is to show how nudging and rational persuasion can complement each other.

The set-up of this chapter is as follows. In section 3.2, I will argue that rational persuasion is characterized by the *aim* of using both arguments and 'non-argumentative means of persuasion' so as to allow and empower individuals to change their attitudes and behaviour based on largely unbiased reflection on arguments.[23] From this characterization, I derive three constraints on the application of psychological knowledge to the design of non-argumentative means of persuasion. To be fit for an attempt at rational persuasion, non-argumentative means of persuasion should not i) bypass, ii) inhibit, or iii) bias recipient reflection on arguments.

In section 3.3, I will use my characterization of rational persuasion to analyse four examples of nudging (labels with carbon footprints, use of defaults, the Ambient Orb, and organ donation campaigns using social norms (Thaler & Sunstein: 2009, 8–9, 190–192, 203–205, 206)). This will reveal that the nudg-

[22] To be fair, Hausman and Welch acknowledge that "actual persuasion is rarely purely rational" (Hausman & Welch: 2010, 135). However, for a meaningful normative comparison between rational persuasion and nudging, we need a characterization of rational persuasion that *takes into account* the facts of behavioral economics. It is unhelpful to define rational persuasion as an ideal that, due to our actual psychological make-up, will seldom be reached. We need to know which use of psychological knowledge is compatible with the aim of rational persuasion, viz. argument-based attitude change.

[23] I do not mean to claim that heuristic processing is necessarily inferior, as it can contribute to being rational and also, some people's cognitive styles may render them more dependent on reliable heuristics (cf. Todd & Gigerenzer: 2000). Crucially, unbiased reflection should remain possible for motivated recipients. Also, cues that trigger heuristics should not lead recipients that are less motivated to reflect on arguments to make decisions that they would not, upon reflection, endorse.

ing approach is relatively broad,[24] and a few instances are even very similar to rational persuasion. Most instances, however, are very different from rational persuasion because they are designed to 'exploit', in one way or another, our unreflective, automatic cognitive processes. Thus, one very important difference between rational persuasion and 'paradigmatic' nudging will become evident: the degree of transparency. In cases of rational persuasion, it is crystal clear that an attempt at influencing is going on and recipients are in a good position to govern their response appropriately.[25]

Nevertheless, in section 3.4, I will argue that we need not choose between rational persuasion and nudging because they have the potential to complement each other. By combining rational persuasion and nudging, governments may be able to combine respect for autonomy with effective behaviour change. Rational persuasion should make citizens realize that they can better adapt their preference and change their behaviour. If this occurs, citizens often will welcome nudges that support them in behaving differently.

An analysis of the differences between nudging and rational persuasion, together with how they may be combined, is of interest to better understand persuasive technology. Despite the label 'persuasive', most persuasive technologies have more similarities to nudging than to rational persuasion. Several psychological insights into influencing human decision-making are employed in both nudging and persuasive technology. Often persuasive technologies are actually 'behaviour change support systems' (Oinas-Kukkonen: 2010) that help users who already have a goal to change their behaviour (see also 4.3.5 below). Thus, as is the case with nudging, rational persuasion remains essential both to inform citizens and to convince them of the need for behaviour change.

I should stress at the outset that the main focus of this chapter is conceptual. I aim to characterize the concept of rational persuasion in a way that is faithful to how it is generally used. I do not claim that it is the only justifiable means for governments to influence citizens. However, I agree with rational persuasion

[24] Several authors make this observation and state that Thaler and Sunstein do not give a coherent definition of what counts as a nudge (cf. Bovens: 2009; Grüne-Yanoff & Hertwig: 2016; Hausman & Welch: 2010; Saghai: 2013a)

[25] Cf. the discussion on having and exercising 'attention-bringing capacities' to detect nudges in (Dworkin: 2013; Saghai: 2013b). In case of rational persuasion these capacities are much more likely to be engaged in order to detect ways in which one does not want to be influenced. If the effectiveness of nudges *depends* on these capacities not being triggered even if they could have been, then these capacities do not deliver the relevant kind of control needed for autonomy.

proponents that in general, compared to other influence methods, rational persuasion maximally respects citizen autonomy. Likewise, I also do not claim that nudging is only justifiable if preceded by rational persuasion that leads to a citizen's informed consent to the nudge. I merely claim that a combination of rational persuasion and nudging escapes the charge of manipulation while having a substantial potential for influencing citizens.

3.2. An account of rational persuasion

Although rational persuasion aims at argument-based attitude change, it also includes limited use of 'non-argumentative means of persuasion' (such as emotions, persuader credibility, and expressing consensus). This notion of rational persuasion is consistent with the widespread linguistic use of the term, in which arguments play a central role. Here, 'argument' can be described as a proposition that confers (logical or evidential) support to another proposition. Persuasion is labelled rational because attitude change is the result of processing arguments (Benn: 1967; Blair: 2012). The underlying idea is that there is a relation of genuine support between the arguments and the position advocated and adopted (Goldman: 1999).[26] The provision of arguments is an excellent way

[26] The term 'rational' in rational persuasion has both a more descriptive and a more normative or evaluative meaning. Moreover, these meanings are interrelated. 'Rational' in the descriptive sense of rational persuasion, refers to the *process* of persuasion that involves the offering and processing of arguments for the advocated position. 'Rational' in the evaluative/normative sense refers to whether or not the persuaded person changes her views on the basis of adequate grounds. Thus, it is rational to be persuaded to views that are supported by a sufficiently wide range of reasons or considerations. It seems plausible that the descriptive sense is derived from the evaluative/normative sense: because it is generally rational (evaluative sense) to change one's views on the basis of good arguments, an attempt at argument-based persuasion is often labeled 'rational persuasion' (descriptive sense).
In this chapter, my first and main focus is on rational persuasion in the descriptive sense. I perform a conceptual analysis in order to understand what persuaders should and should not do so as to allow and empower individuals to change their attitudes and behaviour based on largely unbiased reflection on arguments. Generally, if individuals indeed change their views in such a way, this change will be rational (evaluative sense). However, for purposes of this thesis' research, my main evaluative or normative interest is not in the rationality of the outcome of rational persuasion, but rather in its impact on autonomy. Compared to other influence methods rational persuasion as argument-based change is an influence method that confers to citizens a high degree of control over their attitudes and behavior. Rational persuasion is therefore an outstanding method to respect citizen's autonomy.

to engage a person's deliberative capacities and thus to respect and to support personal autonomy or self-governance.[27] For this reason, rational persuasion is widely regarded as the moral high road in attempting to influence others. However, as I will explain below, safeguarding this possibility of argument-based attitude change requires clear constraints on the use of non-argumentative means of persuasion.

Psychological theories of persuasion confirm that argument-based attitude change is possible. However, they also reveal a large potential for non-argumentative means of persuasion, especially when humans process the arguments only shallowly. The key idea underlying prominent theories of persuasion,[28] or attitude change,[29] is the notion that humans are 'cognitive misers' (Fiske & Taylor: 1984). We simply lack the mental capacity to scrutinize each of the many persuasive attempts we encounter on a daily basis. Therefore, we only make a cognitive effort if we are motivated and have sufficient abilities. In terms of the Heuristic Systematic Model of persuasion, in such cases we engage in effortful and slow analytic or systematic thinking about the persuasive message (Todorov, Chaiken, & Henderson: 2002). We focus on the arguments and other relevant information, and if we change our attitude, we do so based on our reflection. Persuasion research has shown that under conditions of effortful processing, stronger arguments lead to more attitude change (see e.g. Petty,

It is instructive to note that while rational persuasion is an excellent way to help citizens to rationally (evaluative sense) change their mind, it does not follow that it is always irrational for citizens to do so on a basis other than arguments. For example, if one has little time to make a decision and little depends on making a good decision, it can be perfectly rational to follow the majority. For further discussion of the relation between rational persuasion and rationality, see Blair (2012).

[27] See 2.3.1. above for a characterization of 'personal autonomy'.

[28] Prominent psychological theories of persuasion belong to the same family as the dual process theories of human judgment and decision-making used in behavioral economics (see e.g. Kahneman: 2003). As such these two sets of theories share quite some characteristics, including an important place for the role of heuristics. Here I will focus on psychological theories of persuasion as providing the most detailed empirical knowledge of persuasion, while they fit the broad perspective on human cognition of behavioral economics.

[29] In psychological theorizing, persuasion is taken as equivalent to attitude change, which is the focus of research. Here the psychological concept of attitude can be characterized as the evaluation of persons, objects or states of affairs, which has a cognitive, affective, and action guiding component to it (cf. Ajzen & Fishbein: 2000; O'Keefe: 2002). However, researchers would not deny that persuasion may also target other mental states that play a role in determining behavior.

Cacioppo, Strathman, & Priester: 2005). However, if we lack motivation, ability, or both, we will engage in non-effortful 'heuristic processing' of which we are not always aware. This non-analytic mode uses simple cues as its input. Here, a cue can be described as "that subset of information that enables [the use of] simple decision rules or heuristics to form a judgment" (Todorov, Chaiken, & Henderson: 2002, 197).[30]

Heuristic or non-systematic processing of attempts at persuasion may yield accurate results but may also lead to well-known biases (cf. Tversky & Kahneman: 1974). Much depends on the fit between heuristics and the cues provided by the cognitive environment. Take, for example, the well-researched cue of 'persuader expertise'. Under conditions of non-systematic processing, perceived expertise functions as a cue that leads to increased persuasion, measured in terms of increased attitude change (Chaiken & Maheswaran: 1994; Petty, Cacioppo, & Goldman: 1981). Presumably, some heuristic, such as "expert sources express credible opinions", is at work (O'Keefe: 2002). If the persuader is rightly perceived as an expert or credible, then there is a fit between the cue and heuristic processing such that it may very well be rational to (unreflectively) adjust one's attitude. If, on the other hand, the expert merely *appears* to be an expert but is not, heuristic processing will lead to bias and unwarranted attitude change.

Since an account of rational persuasion should clarify which *intentional uses* of psychological knowledge to influence are acceptable, I employ the notion of 'non-argumentative means of persuasion'.[31] This notion is not equivalent to the notion of cues: not all cues present in a persuasion context are deliberately introduced and employed as a *means* to persuade. Conversely, non-argumentative means of persuasion do not always function as cues in heuristic processing. Thus, not all cues are non-argumentative means of persuasion, and not all non-argumentative means of persuasion are cues.

[30] Another prominent theory, the Elaboration Likelihood Model (ELM) (see e.g. Petty, Cacioppo, Strathman, & Priester: 2005), makes a similar distinction between two modes of processing persuasive messages, though it stresses a continuum of processing intensity. In addition, the model allows for more types of non-systematic processing than heuristic processing. Nonetheless, the commonalities between the models are more important than their differences and many empirical findings support both models.

[31] I am inspired by Blumenthal-Barby's use of the phrase "non-argumentative influence" (Blumenthal-Barby: 2014), though the meaning of her phrase only partly overlaps with the meaning of this chapter's "non-argumentative means of persuasion".

It is a central insight of persuasion research that one and the same means of persuasion can impact persuasion in different ways, depending on how much cognitive effort the recipient invests (Todorov et al.: 2002). Another prominent theory of persuasion, the Elaboration Likelihood Model, distinguishes as many as five such different ways, or so-called processes (Petty & Briñol: 2008a). Although my account of rational persuasion is not dependent on a precise identification of all possible psychological processes, it is instructive to illustrate these five with emotions (Petty & Briñol: 2008b; Petty, Schumann, Richman, & Strathman: 1993). Emotions form an apt example since they have been an important non-argumentative means of persuasion throughout the history of rhetoric and persuasion. For example, in his *Rhetoric*, Aristotle recognized that "[w]e do not give judgment in the same way when aggrieved and when pleased, in sympathy and in revulsion" (Aristotle: 1991, Bk 1.2, 1356a). In addition, emotions are employed in some nudges, which is a further reason to discuss emotions as an example of how non-argumentative means can shape persuasion.

To start with the first of five ways in which emotions can impact persuasion, emotions can bypass recipient reflection and serve as a cue for heuristic processing. This is most likely under conditions in which a recipient is neither motivated nor able to engage in effortful reflection on the persuasive message. Via a low-thought process such as the 'affect-heuristic' (Zajonc: 2001), one's emotional state or mood becomes associated with the persuasive issue. Thus, if one is happy, the affect heuristic will lead to more positive attitudes towards the issue than if one is in a more neutral mood. Second, emotions affect the depth of processing under conditions where this depth is not yet determined by factors as ability and motivation. Generally, happiness seems to decrease and sadness seems to increase message processing. Third, under conditions of effortful reflection on a persuasive message, emotions affect the valence of thinking. One way in which emotions have this impact is by increasing the likelihood that thoughts with a corresponding valence will be generated or retrieved. Thus, for example, when a sad person reflects on the possible consequences of the advocated plan to build a new nuclear site in her neighbourhood, negative consequences will more easily come to her mind. Evidently, then, emotions may have a biasing effect and distort sound thinking (for a brief discussion of the notion of 'biasing', see below). Fourth, under conditions of effortful processing, emotions can themselves be the *object* of reflection. Recipients may wonder how to interpret their emotion in relation to the persuasive issue. For example, they may ask: 'Is my fear justified by the risky cancer treatment the doctor is attempting to persuade me to choose?'. Fifth and finally, emotions can affect the recipient's confidence in her thoughts developed in response to a persuasive message. The greater this confidence, the more impact these thoughts have on attitudes. Generally, happiness has a positive effect on confidence and sadness has a negative effect. How these emotions ultimately affect attitudes depends on the content of the after-message thoughts. Happiness will increase the impact of both positive and negative thoughts by increasing the recipient's confidence in them. In this overview of the five processes, emotions merely served as an illustration for how any means of persuasion can impact attitude change. Thus, in principle, similar analyses could be made for other non-argumentative means, given the relevant empirical research has been done with respect to that means.

This empirical perspective confers substantial insight into the various roles that can be played by non-argumentative means of persuasion. In the present example, it allows for a nuanced view on the role of emotions in rational persuasion, differentiating between potentially positive and negative roles. It helps transcend one-sided oppositions between reason and emotions and can acknowledge that emotions are sometimes a source of knowledge (cf. de Sousa: 2014; Roeser: 2006).

However, this empirical perspective also reveals a complication for the project of characterizing rational persuasion. That is, it becomes clear that it cannot be said in general whether the use of a particular non-argumentative means of persuasion is compatible with aiming to arrive at argument-based attitude change. We saw above that emotions can stimulate reflection, which is conducive to the aim of rational persuasion. However, they may also, at the same moment, bias reflection by biasing the valence of a recipient's after-message thoughts (or by biasing confidence in these thoughts). In addition, it may not be clear whether an emotion experienced by the recipient is skilfully induced by the persuader or whether instead it is the recipient's response to the persuasive message. As already noted by Aristotle, this is a relevant difference.[32] Also relevant is the extent to which a recipient can and will be aware of and can counter possible biasing effects. Similar stories can be told for other non-

[32] Aristotle holds that a legitimate appeal to or use of emotions is one that brings the audience in a certain emotional state *through* the argument itself, the *logos* (Aristotle: 1991, Bk 1.2, 1356a).

argumentative means of persuasion and consequently, it is not possible to compile a simple list of means that are acceptable as part of rational persuasion.

To address these complexities, I will develop some constraints on the employment of non-argumentative means in rational persuasion. These constraints follow from the notion of rational persuasion introduced above: rational persuasion is aimed at allowing and enabling attitude change as the result of the recipient's largely unbiased reflection on argument. Broadly speaking, we can conceive of three ways in which an attempt to influence can fail to qualify as rational persuasion: it can bypass, inhibit, or bias recipient reflection on arguments. The constraints on the use of non-argumentative means of persuasion correspond to these three ways. First, to avoid bypassing, a persuader should not include one or more non-argumentative means of persuasion so as to (also) bring about attitude change under conditions of heuristic processing. Indeed, he should go to some length to prevent such attitude change, because it would not be based on arguments. The point is not in the first place that designing non-argumentative means of persuasion for heuristic processing will often count as manipulation. Rather, doing so is simply inconsistent with the goal of rational persuasion. Second, non-argumentative means of persuasion should not inhibit reflection on the arguments, and ideally, they stimulate such reflection. Thus, for example, it is incompatible with rational persuasion when persuaders try to prevent people from thinking about the arguments by making them happy, telling them that their view is the consensus view and carefully devising an air of expertise and authority. Each of these three non-argumentative means of persuasion inhibits reflection. Third, when recipients actually do engage in effortful processing of the arguments, non-argumentative means should not bias this reflection. Reflection can be biased in many ways, as has become clear in the above discussion of the processes that underlie persuasion. Here, bias refers to any deviation from sound reasoning. Roughly speaking, reasoning qualifies as sound when it draws logically correct inferences from a sufficiently wide and balanced range of facts and considerations.[33]

[33] Whereas these constraints are inspired by the Heuristic Systematic Model and the Elaboration Likelihood Model, they are in fact suggested by the dual-process approach to cognition in general. For, all the different versions of dual process have a distinction similar to that between analytic processing and heuristic processing. Furthermore, all identify factors that determine which type of processing will dominate (and thus whether there will be bypassing or inhibition of reflection on arguments), and they all work with

The three constraints need to be applied in combination: each may limit or qualify the application of the other. To see how, consider again 'persuader expertise'. Let us stipulate that the persuader is indeed an expert on the topic. Under conditions of heuristic processing, perceived persuader expertise will lead to more favourable attitudes towards the persuasive issue. In principle, this would be ruled out by the first constraint. However, if a persuader manages to let the audience carefully think about the arguments, recipients will be more confident about their thoughts developed in response to the message. Consequently, in that case, persuader expertise fosters argument-based attitude change, which is a reason to limit application of the first constraint. Moreover, since it is rational to have more confidence in thoughts developed in response to an expert message, this non-argumentative means of persuasion does not cause bias and the third constraint is satisfied. Indeed, recipients would be disadvantaged if they lacked an indication of the persuader's expertise. A related reason not to constrain the role of persuader expertise involves fairness. Even though the audience may engage in non-argument-based attitude change under conditions of heuristic processing, it seems that if one is an expert, one is entitled to receive at least the corresponding benefits under conditions of effortful reflection.

The three constraints must be applied not only in combination but also to the persuasive attempt as a whole, thus including the arguments and the total set of non-argumentative means of persuasion (instead of individual means separately). The main reason involves the importance of stimulating reflection. It is an effect of persuader credibility that if the level of reflection is not yet determined, there might be a decreased likelihood that a recipient will think carefully about the arguments (Petty & Briñol: 2008b, 57). If the constraints were applied to this non-argumentative means of persuasion in isolation, it would be ruled out by the first constraint. However, the persuader may ensure a sufficient level of reflection, for example, by making the issue's relevance salient such that persuader credibility will not threaten the aim of rational persuasion. We see that a complete attempt at rational persuasion can satisfy the three constraints in a way that individual non-argumentative means of persuasion could not.

the notion that processing can be biased in various ways (e.g. J. S. B. T. Evans & Stanovich: 2013; Strack & Deutsch: 2004).

Still, adherence to the three constraints will not always ensure sufficient, completely unbiased reflection on the arguments. Some recipients are difficult to motivate or have a greater than average vulnerability to biased thinking. After all, there are substantial individual differences in the capacity for rational thought (Stanovich: 2010). It is for this reason that I characterize rational persuasion in terms of *allowing for and aiming at* attitude change based on *largely* unbiased reflection on arguments. Inherent in this account is the acknowledgment that even a well-designed attempt at rational persuasion, in which the non-argumentative means of persuasion are carefully chosen, may fail.

There is a more fundamental reason not to demand completely unbiased reflection. 'Bias' is a normative notion, which may be difficult to 'cash out' in persuasive contexts in a manner that is uncontroversial. For many of the decision-making problems studied in behavioural economics, the normatively correct answer is relatively uncontroversial, at least if statistical reasoning provides the standards (but for discussion, see (Samuels, Stich, & Bishop: 2002)). For example, with the use of anchoring heuristics, it will often be possible to assess whether or not such a heuristic operates on a useful piece of background knowledge, leading to a statistically correct estimation (Tversky & Kahneman: 1974). Persuasive issues are typically more complex. Determining which set of background knowledge is sufficient for unbiased reflection can be difficult and can itself be a matter of dispute. Another way of articulating this contrast is to note that in the heuristics and biases literature, it is mainly the notion of theoretical rationality that is at stake in distinguishing biased from normatively correct reasoning. In persuasion contexts, it is most often the wider and more complex notion of practical rationality. With respect to issues such as tax policies and other controversial topics, we often hold that reasonable people may disagree.

As a result, my account does not allow for an easy and straightforward identification of instances of rational persuasion. Instead, application of the constraints entailed by the account will often require empirically informed, case-by-case judgment. This is no weakness of my account, however; given the subject matter, it is inevitable. It still compares well with existing definitions of rational persuasion, which are either uninformative or of little practical guidance *as such*. For example, Hausman and Welch define rational persuasion as the attempt "to persuade by means of fact and valid arguments" (Hausman & Welch: 2010). Burnell and Reeve define their "central notion of persuasion" as "*A* gets *B* to do / believe / accept / reject something which he would not otherwise do / believe / accept / reject, by exhibiting reasons or consequences of alternatives confronting *B*" (Burnell & Reeve: 1984, 409–410). They "add that *B*'s ratiocination is essential to persuasion". Beauchamp and Childress hold that "…a person must come to believe in something through the merit of reasons another person advances" (Beauchamp & Childress: 2009).[34] My account shares with these definitions the core idea of persuasion as argument-based change of attitudes or beliefs.

However, the weakness of these characterizations is that they are silent on the role of non-argumentative means of persuasion, sometimes explicitly denying them any legitimate role (e.g., Benn: 1967). Consequently, they are of no help in deciding which non-argumentative means of persuasion to include so as to reach the aim of rational persuasion. Likewise, they do not indicate the cases in which a non-argumentative means of persuasion should be excluded. Thus, the cited characterizations of rational persuasion fail to recognize the positive role of non-argumentative means of persuasion, and they fail to guide the resolution of recurrent disagreements about which non-argumentative means fit the aim of rational persuasion and which do not. The account developed above has the resources to do better in both respects.

To conclude this section, it was shown that recent decades' empirical findings on human cognition do not question the viability of rational persuasion as an approach to influencing citizens. Indeed, psychological theorizing on persuasion shows several ways in which to foster reflection on arguments. Therefore, even though choice architects inevitably set up decision-making contexts in *some* way, they still can engage citizens' deliberative capacities by providing them with facts and arguments. Nudging is not the only option.

3.3. Rational persuasion and nudging contrasted: the account applied to nudging

Now that we have a better view both of what rational persuasion is and of the role of non-argumentative means of persuasion, it is fruitful to contrast rational persuasion and nudging. In this section, I will use my characterization of rational persuasion to analyze four nudges. I aim to clarify some of the concep-

[34] The authors give their definition in the course of summing up 'forms of influence' which are relevant in medical-ethical contexts, so they are not in the business of contributing to conceptual clarification of persuasion.

tual differences between rational persuasion and various types of nudging. (As noted above, nudging is a relatively fuzzy concept, and several different types of actions are labelled as a 'nudge' although they are different with regard to their underlying mechanisms and how they influence humans.) In general, a nudge could be very similar to rational persuasion, but most are quite different. However, although rational persuasion and nudging are different, it might be that a nudge could act as a non-argumentative means of persuasion in rational persuasion. Thus, even if a nudge may not qualify as rational persuasion, it might be part of an attempt at rational persuasion. First, I start with a (perhaps atypical) instance of nudging that qualifies as rational persuasion.

Labels that displays the carbon footprint of consumer goods

Thaler and Sunstein discuss the potential of labelling as a means to raise consumers' awareness of their impact on global warming. Several countries are implementing initiatives to display products' carbon footprints. Consumers are made aware of the fact that buying goods increases global warming caused by CO_2 emissions. When this information is imparted on a product label, consumers have the ability to choose a lower-emission alternative. Large and abstract problems are translated into small consumer decisions.

Given that a consumer knows the relevance of emissions and is sufficiently motivated, numbers provide an argument that is relevant to the decision of which product to use. Therefore, labels either can be part of attempt at rational persuasion or indeed, can be such an attempt in their own right.

Organ donation campaigns that employ social norms

Thaler and Sunstein's example of organ donation campaigns that involve appeal to social norms has some similarity to rational persuasion. However, as I will argue, it is relatively uncertain whether social norms can contribute to argument-based change, and in any event, this might not be the aim of the campaign. Thaler and Sunstein describe the nudge as follows. A website makes arguments that emphasize the importance of having enough donors. In addition, it states that "87 of adults in Illinois feel that registering as an organ donor is the right thing to do", and "60 % of adults in Illinois are registered organ and tissue donor". They explain the effectiveness of this nudge by stating that "people like to do what most people think it is right to do" and "people like to do what most people actually do" (Thaler & Sunstein: 2009, 184–192).

This explanation refers to the Focus Theory of Normative Conduct, developed by Cialdini and colleagues (Cialdini et al.: 2006; Kallgren, Reno, & Cialdini: 2000). This theory "asserts that norms are only likely to influence behaviour directly when they are focal in attention and, thereby salient in consciousness" (Cialdini et al.: 2006, 4). Furthermore, the theory distinguishes between two types of social norms. Descriptive norms "refer to what is commonly done", and motivate, according to the theory, by "providing evidence of what is likely to be effective and adaptive action". Injunctive norms "refer to what is commonly approved or disapproved", and "motivate by promising social rewards and punishments" (ibid., 4). Illinois citizens who are not registered donors deviate from both the descriptive and the injunctive norms elicited by the campaign. The mechanisms just mentioned could result in the citizens' registration.

Assuming the validity of Cialdini's Focus Theory, my account of rational persuasion is helpful to elucidate the extent to which the Illinois campaign—taken as a whole—qualifies as rational persuasion. More generally, the question is whether appeal to social norms could fit the aim of rational persuasion and, thus, whether social norms can be legitimate non-argumentative means of persuasion. First, however, I need to clarify what the theory means by social norms being 'in focus'. It turns out that this does *not* mean effortful and systematic processing of the pros and cons of a certain course of action in which a person deliberates on how the social norm should affect her behaviour. Instead, the Focus Theory is based on studies in which the social norms are activated in participants' minds, e.g., by reading a diary fragment and thus, by priming them. Subsequently, the participants were required to perform an unrelated task in which they could show norm-following behaviour (Kallgren et al.: 2000). Clearly, these participants could merely have been acting without engaging in an effortful and conscious deliberation about how the social norm should affect their behaviour.[35] This seems likely to be the case.

Now I apply the three constraints developed in the previous section to the role of social norms in the Illinois campaign. Given the above analysis, under conditions of heuristic processing, the descriptive and injunctive social norms will lead to (some) attitude change among non-registered citizens. They may employ a heuristic such as 'often, the majority acts rightly', or they may be

[35] This interpretation is confirmed by Cialdini's likening his Focus Theory with heuristic models of cognition (Kallgren, Reno, & Cialdini: 2000) and by his general emphasis on the automatic nature of human responses to many attempts at social influence (Cialdini: 2006).

driven by an implicit fear of failing to conform to the majority norm. In Thaler and Sunstein's interpretation, this impact of social norms was intended by the designers of the Illinois campaign. If correct, argument-based change was not the aim, and the first constraint is not satisfied. This is also confirmed by the website's headline and recurring slogan "I am. Are you?". With regard to the second constraint, social norms may very well inhibit reflection on arguments. In the campaign, both the injunctive and the descriptive social norms tell a citizen that the advocated position—i.e., signing the donor form—is the position held by most other citizens. Persuasion research has shown that under conditions where the level of reflection is not yet determined, this type of 'source majority' most often leads to less effortful elaboration of the arguments than source minority (Horcajo, Petty, & Briñol: 2010). Most likely, therefore, the second constraint is not satisfied. With regard to the third constraint, it is difficult to assess whether the social norms in the campaign will bias citizens who think carefully about whether or not to become donors. Persuasion research indicates that a majority source leads recipients to develop more positive thoughts in response to the message, leading to increased attitude change (Horcajo, et al.: 2010; Martin, Hewstone, & Martin: 2007). Thus, if citizens think about the current lack of sufficient donors and the tremendous benefits to those receiving an organ, and develop positive thoughts in response, psychological theory predicts that the social norms stated in the campaign will make their thoughts even more positive. I will not attempt to assess whether or not this impact of social norms counts as bias. In some cases, it might be practically rational to be guided by the majority view on what is the right thing to do; in others, it will lead us astray. However, to know which cases are which, we need to know whether or not the advocated position is the right one. Therefore, disagreement about whether presenting social norms leads to biased reflection reduces to disagreement about the rightness of the advocated position itself. Even if being influenced by majority norms usually supports being practically rational, it does not justify the use of social norms in individual attempts at rational persuasion.

The result is that the Illinois campaign does not qualify as an attempt at rational persuasion. It does not appear intended to foster reflection on arguments and argument-based attitude change. It seems to be designed for mere effectiveness in terms of more registered donors. However, people may have many reasons—whether valid or not—for not signing the form. People may be reluctant to consider becoming donor out of a fear of death, or they may be afraid that doctors will not give them fully adequate medical treatment if they are registered donors (Nijkamp, Hollestelle, Zeegers, van den Borne, & Reubsaet: 2008). Instead of bypassing such worries by employing the power of social norms, rational persuasion involves providing arguments relevant to these worries. Regarding the use of social norms as a non-argumentative means of persuasion in general, the above discussion shows that rational persuaders must be careful. In some cases, appeal to social norms can increase reflection. For example, when the position advocated is supported by a minority, recipients will devote more effort to reflecting on the arguments. However, it is likely that as a consequence, those recipients will develop less positive thoughts. Social norms might easily lead to biased reflection, making them less fit for the aim of rational persuasion.

Use of defaults

In many choice settings, making one option the default has a considerable impact on what people choose. Thaler and Sunstein's central examples involve pension saving plans (Thaler & Sunstein: 2009, 113–128). If employees have to actively enrol in a retirement saving plan, many either do not manage to take the time to do so or simply forget to do so. If, however, enrolment is the default setting, and employees are automatically enrolled, participation rates drastically increase, with few participants opting out.

Thaler and Sunstein explain the impact of defaults in terms of the status quo bias. For several reasons, humans are often slow to change their situations. In some cases, this tendency involves loss aversion, e.g., when increasing one's rate of savings would be wise and possible, but one simply does not want to receive a lower salary. Other cases involve a mere lack of attention. Sometimes, if people decide that they want to join a savings plan or increase savings rates, they do not manage to make the small effort of completing the form (inertia). The idea underlying nudging through a clever default setting is to avoid or even to employ these psychological phenomena. For example, if enrolment is the default, then inertia works in favour of employees, based on the assumption that they indeed would wish to join.

We can identify different ways of nudging by default settings, which gradually come closer to rational persuasion. The difference between nudging and rational persuasion is the greatest when choice architects set a default that leads to a change in someone's situation, e.g., enrolment in a non-mandatory savings plan, without giving notice. This amounts to deciding for the employee that it would be better for him to join. Adding notification by writing the person in question a letter that informs him about enrolment, however, at least provides the person with an opportunity to consider whether or not he wants to participate. However, this still does not provide him with arguments in favour of continued participation. Therefore, the next step would be to add to the notice of participation an explanation of the basic characteristics of the savings plan and why it is regarded as a good choice for a specific type of employee. This explanation should also describe how to opt out of the plan (which should as easy as possible). This accompanying information makes the default setting maximally transparent. It provides the person with an opportunity to assess whether he is sufficiently similar to the envisioned employee and whether the reasons for joining apply in his case.

This last method of nudging by setting defaults is the closest to rational persuasion. It is an open attempt at influencing choices, and arguments are made in favour of the target behaviour, viz. accepting (or rather, not rejecting) the default. Subjects can reflect on the arguments and decide whether they indeed wish to accept the default. It is a bit of a stretch to view the default setting as a non-argumentative means of persuasion, but it remains instructive to apply the constraints. Regarding the first constraint, it is possible to choose a default that is believed to make most employees better off while having the goal that each employee makes her own choice about enrolling in a savings plan. Second, the default can both stimulate and inhibit reflection on the arguments. Some employees will be happy not to worry about savings plans, which they might perceive as complicated. As explained above, additional non-argumentative means of persuasion, such as making personal relevance salient, may be added to stimulate reflection. Others will dislike the default setting and check carefully whether the plan indeed fits them. Third, the default setting as such may bias the employee's careful assessment of the explanation in favour of the default savings plan. Those employees who dislike paternalism may be overcritical (cf. Anderson: 2010), whereas others may attach too much value to default setters' claim that the savings plan benefits them.

Even though nudging by means of default setting can to some extent be complemented with rational persuasion, the crucial point is the direction in which inertia works. If rational persuasion fails, people do not enrol in a savings plan; if it succeeds, they still must overcome inertia to join the plan. In contrast, if people are not persuaded by the arguments for the default, they still must take active steps to opt out. Depending on how easy this is, inertia will represent a barrier. In conclusion, defaults accompanied with arguments have some similarity with rational persuasion, but the fact that a default setting can change one's situation without one's being persuaded is a crucial difference.

At this point, it is instructive to briefly discuss a third possible explanation of the effectiveness of default setting. In addition to loss aversion and inertia, Güne-Yanoff and Hertwig (2016, 2016), discuss the possibility that defaults function as a signalling device. Employees interpret the default as the option chosen by the designers of the retirement savings plans as best for them. Thus, defaults have the effect of giving the employees information and recommending one of the options. The authors note that there are even more possible mechanisms that explain how defaults work. Also, different mechanisms may be at work for different employees, or even for one employee. Of course, this fact complicates the analysis given above. If we briefly go through the different ways of setting defaults again, first, it is clear that setting enrolment as the default without giving notice prevents the default from functioning as a signal, since the employee is not aware of the default at all. Merely giving notice of enrolment already enables the signalling function of the default. However, giving notice of enrolment is still not yet giving arguments, even if the default can rightly be interpreted as the designer's sincere and informed judgment of the best option for employees. If the default is accompanied by an explicit explanation, then this overlaps with the signalling function, which has no independent impact any more, unlike inertia that still has to be overcome by an employee who concludes that he prefers to opt out. From these observation, it follows that the comparison between rational persuasion and nudging is not straightforward due to uncertainty with respect to the mechanisms at work in a specific nudge (cf. Grüne-Yanoff: 2016). However, we can safely conclude that the differences between nudging by setting defaults and rational persuasion are significant.

The Ambient Orb (a glowing ball providing ambient feedback on energy use)
One type of nudging involves an ingenious way of giving feedback. Obtaining the right kind of feedback on actual behaviour is an important determinant of successful behaviour change. The Ambient Orb is a device that provides direct and simple feedback on energy use. This little ball glows red during peak hours and green otherwise, leading users to significantly reduce their energy consump-

tion.[36] Thaler & Sunstein argue that the Orb is so effective because "it makes energy use visible" (op cit., 206).

Figure 3.1 The Ambient Orb[37]

However, the psychological mechanisms underlying the Orb's effectiveness are not easy to identify. According to Thaler & Sunstein, the "flashing red ball really gets people's attention and makes them want to use less energy" (op cit., 206). This is naturally read as expressing the idea that people focus on the changing colour of the ball, deliberate, and form an intention to change their behaviour to decrease their energy consumption. Nevertheless, the core idea underlying ambient feedback is to provide a form of feedback that does *not* require focal attention, but which is nonetheless processed in a way that leads to behaviour change.[38]

[36] Thaler and Sunstein refer to http://archive.wired.com/techbiz/people/magazine/15-08/st_thompson.

[37] http://ambientdevices.myshopify.com/products/energy-orb (accessed 19-3-2015)

[38] Thaler and Sunstein describe the psychological mechanism by means of which the orb causes humans to reduce energy use too much in terms of analytical processing: the "flashing red ball really *gets people's attention* and makes them *want* to use less energy" (206, emphasis added). This suggests that users are thinking about their energy use, have an explicit and at that moment conscious preference to reduce energy use, and then actively decide to adapt and use less. But the whole idea of ambient feedback is to employ cognitive processes that do *not* require focal attention and explicit desires. For an instructive picture see http://www.ambientdevices.com/about/the-science.

In addition, instead of *generating* the wish to reduce energy consumption, the Ambient Orb seems to depend on a pre-existing motivation or goal, to which the feedback is targeted. Users must decide to place the Orb on their desk, and they know what the changing colours mean relative to their attitudes regarding energy consumption. They also know what kinds of behaviour lead to decreased energy use. Without identifying the relevant cognitive mechanisms, we might say that ambient feedback 'triggers' the relevant behaviours in a way that does not demand explicit attention and deliberation.

This analysis of the Orb is supported by research into the effect of different types of feedback on energy use. Ham and Midden found that in an experimental task, lighting feedback resulted in significantly less energy use than factual feedback (Ham & Midden: 2010). In their experiment, users were explicitly given the goal of saving energy. Arguing from the effect of their second variable, cognitive load, Ham and Midden explain this result in terms of different demands on cognitive resources. Processing actual feedback in the form of kWh-consumption figures requires focal attention and involves comparison with certain standard values. Lighting feedback is easier to process since it requires no focal attention and the meaning of the colour is immediately evident, requiring no effortful comparison with a standard.

In addition to the ease of processing lighting feedback, there might be further explanations for the effectiveness of the Ambient Orb. Ham and Midden suggest that the lighting feedback may either affect a user's mood or function as social feedback by eliciting the social norm. A similar suggestion made in relation to the Orb points to the cultural meaning of the colours green and red, signalling approval and disapproval (Selinger & Whyte: 2011). Further research should provide more clarity about the relative contribution of these various factors to the effectiveness of lighting feedback.

As a consequence of this uncertainty about the cognitive mechanisms underlying the Orb's influence potential, it is not at all straightforward how this type of nudge compares to rational persuasion. Nevertheless, one difference is clear: attempts at rational persuasion aim at the recipient's effortful reflection on the message and its arguments, whereas the Orb is explicitly designed to be effective *without* such reflection. Corresponding to this contrast, rational persuasion often aims at changing the relevant attitudes on the basis of arguments, whereas the Orb seems to rely on pre-existing attitudes favourable to energy conservation. Thus, if the aim is that people reduce their energy consumption, rational persuasion and nudging each target a different mental state. Below, I will elaborate on this idea that the two are complementary means of influence rather than rivals.

Like defaults, Orbs do not function as a non-argumentative means of persuasion in an attempt at rational persuasion. Nevertheless, if viewed as a non-argumentative means of persuasion, the Orb would clearly fail to satisfy the first and second constraint. For, it is designed to change behaviour *absent* user reflection. In case, however, that a user is triggered to think about her energy consuming behaviour, would the Orb's feedback have a biasing impact? Here similar considerations apply as above in the case of social norms regarding donor registration.

For a more determinate evaluation of the Orb as a self-standing influence (not combined with rational persuasion), we need to know whether the Orb will also be effective if users are not motivated to conserve energy, instead preferring comfort. If the Orb would merely be a self-supporting tool to act in accordance with one's attitudes or even explicit goals, then worries about manipulation and substituting the user's judgment and decision-making for the nudger's judgment would dissolve. However, it might very well be the case that if the Orb also influences via affecting moods or by expressing social disapproval, it would do so absent user motivation for energy reduction. Instead, the Orb might make users feel good because of the green colour or induce them to consume less energy to avoid bad feelings caused by the red colour.[39] If this is indeed the case, then there is a concern about manipulation since users may be either unaware or insufficiently aware of why they change their behaviour. As long as we do not know with more certainty why the Orb is so effective, we must suspend judgment. Here, more research is indispensable.

These four analyses of nudging allow some more general (but tentative) conclusions. Most fundamentally, rational persuasion grants citizens more control over their attitudes, preferences, and behaviour than nudging does. Rational persuasion is maximally transparent about the fact that an influence attempt is occurring. In addition, non-argumentative means of persuasion are limited to

[39] Here the question about the accuracy of the feedback becomes an important issue (cf. Spahn: 2011). Who determines the values at which the Orb becomes red? Is using energy during peak hours something that *deserves* social disapproval? Could a purple color, designed to be more neutral, serve as an alternative that saves much of the potential of the Orb? It appears that there are only value-laden answers to the question of the accuracy of the feedback.

means that foster argument-based change. In contrast, several nudges are most effective in the absence of effortful citizen reflection, leading to concerns about manipulation. Fortunately, combining rational persuasion and nudging will often alleviate such concerns while increasing the chances of successful behaviour change. In the next section, I will further develop this last thought.

3.4. Nudging and rational persuasion can be complementary

In the previous section, a picture emerged of nudging and rational persuasion as complementary rather than as rival means to influence citizens. Section 3.2 clarified how, even in light of the insights of the field of behavioural economics, rational persuasion remains a viable means to bring about argument-based attitude change. The previous section showed the potential of nudging as a means to enhance attitude-behaviour consistency. In this section, I will argue how combining rational persuasion and nudging increases the chances for successful behaviour change while safeguarding against morally problematic forms of nudging.

Solving social problems most often requires actual behaviour change. This explains why educational campaigns aiming at rational persuasion are so often disappointingly ineffective in solving social problems (Briñol & Petty: 2006). Mere attitude change may not result in the desired behaviour change for several reasons, as can be explained by Fishbein and Ajzen's well-known 'theory of planned behaviour' (see figure 3.2 below). According to this theory, behavioural intentions and behaviour are determined not only by humans' attitudes but also by the social norms they perceive as relevant to the behaviour and the extent to which they believe they can actually perform the behaviour in question (Ajzen: 2012).

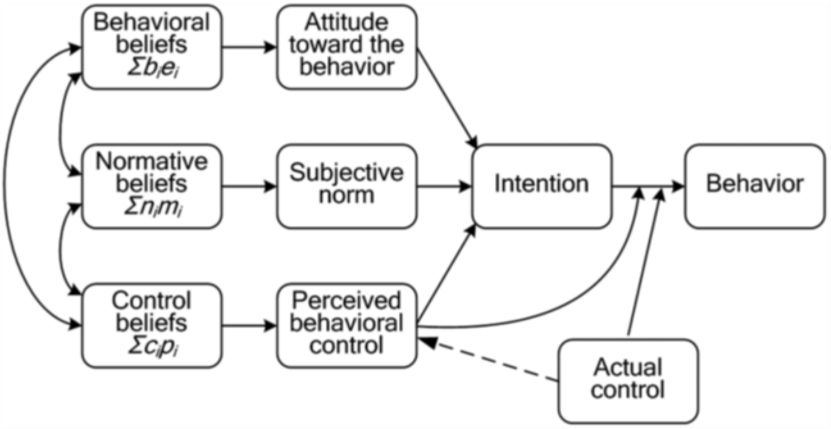
Behavioral beliefs $\Sigma b_i e_i$
Attitude toward the behavior
Normative beliefs $\Sigma n_i m_i$
Subjective norm
Intention
Behavior
Control beliefs $\Sigma c_i p_i$
Perceived behavioral control
Actual control

Figure 3.2 Theory of Planned Behaviour (Ajzen: 2012)

The social issue of reducing energy consumption may serve as an illustration. Education of all sorts provides each citizen with well-known arguments for saving energy. In response to these attempts at rational persuasion, citizens often develop global attitudes in favour of saving energy. However, these are not yet positive attitudes towards specific behaviours aimed at reducing energy use. The theory of planned behaviour holds that only the latter are predictive of actual behaviours, insofar as action, target, context and timeframe are specified. Thus, for example, a positive attitude towards turning off energy-hungry appliances (action) during peak hours (time frame) at home (context) in order to reduce energy consumption (target) is sufficiently specific to make the corresponding behaviour more likely.

But even such specific attitudes will not lead to action if a person does not perceive himself as in control over his actions ('perceived behavioural control'). One might know that one is simply prone not to notice that one's electrical appliances are in use during peak hours, and consequently not develop the intention to switch them off. Or one might have the behavioural intention to switch off the appliances, but simply forget to follow through. In these cases, the Ambient Orb discussed in the previous section will be of tremendous help since its colours straightforwardly indicate peak hours. The Orb is a tool that enables actual control over managing one's appliances and in doing so also strongly increases the user's perceived behavioural control. In this way, the Orb is an excellent nudge for enhancing attitude-behaviour consistency.

Social norms (see also the last section above) are the planned behaviour model's third factor that determines behavioural intentions. Each of us has many beliefs about what the persons and institutions significant to us want us to do. These beliefs combine into a so-called subjective norm that co-determines our intentions. Certainly, many significant others and institutions would like us to reduce our energy use. As suggested above, it might well be the case that the Ambient Orb activates the user's subjective norm that tells her to reduce energy consumption, especially through the use of the symbolic meanings of the colours green and red.

The Ambient Orb example illuminates how rational persuasion and nudging together form a powerful combination for successful behaviour change. Rational persuasion is a means of bringing about attitude change and nudging is a means of helping persons to act on their new attitudes. In addition, combining rational persuasion and nudging helps solve potential moral problems with the latter. In the previous section, I raised the concern that the Ambient Orb might induce users to reduce energy consumption even when they do not have favourable attitudes towards doing so. They could act merely to avoid bad feelings caused by the red colour, perhaps even without realizing it. If so, they are manipulated rather than choosing to reduce energy consumption. If, however, rational persuasion has resulted in favourable attitudes, then users moved to avoid bad feelings would still act in accordance with their attitudes. In addition, as long as users must make an active decision to purchase the Orb, we can assume the relevant attitudes since they buy it explicitly to reduce their energy use. They have made a choice to employ the Orb and can be viewed as nudging themselves towards self-chosen goals. Combining rational persuasion with nudging thus alleviates the worry that nudging involves manipulation. (However, users might not understand the psychological mechanisms underlying the Orb's effectiveness.)

More generally, rational persuasion as a motivator for citizens to engage in self-nudging helps us avoid Thaler and Sunstein's problematic appeal to a certain informed-desire view of individual welfare. According to them, "in some cases individuals make inferior choices in terms of their own welfare – decisions that they would change if they had complete information, unlimited cognitive abilities, and no lack of self-control" (Sunstein & Thaler: 2003, 1162). Many problems have been noted with the notion of welfare tacitly adopted here. First, it creates such a large gap between a person's actual self and her ideal self that she might feel alienated from it. She will likely not acknowledge the authority of that ideal self (Cf. Rosati: 1995; Sobel: 1994). Second, the choice architects designing the nudges are confronted by the same limitations as the citizens they aim to steer into making better choices. They also lack knowledge, have finite cognitive abilities and might lack self-control, which raises the question of how they can know what makes life go well for citizens. Thaler and Sunstein would probably do well to adopt the less controversial idea that in some cases, a competent judge could make reliable judgments that people are making choices that are detrimental to their welfare (Qizilbash: 2012).

However, this would still be no solution for the third problem that even if citizens make bad choices, that fact in itself is, given liberal anti-paternalism, no justification for nudges that amount to manipulation (for a general treatment of liberal anti-paternalism, see Feinberg: 1989). Thaler and Sunstein repeatedly stress that the subjects can opt out if they wish. However, they fail to appreciate that liberty of choice can be interfered with both by blocking options and by subverting or manipulating the process of decision-making. As argued by many authors, several nudges are manipulative because they are designed both to bypass a person's deliberative capacities and to exploit biases and other imperfections in decision-making to ensure that he 'chooses' the alternative preferred by the choice architect (e.g. Grüne-Yanoff: 2012; Hausman & Welch: 2010; Saghai: 2013a). Even if the subject could opt out, a well-designed nudge makes it very unlikely that he actually will do so. For this reason, not all nudges are compatible with liberalism's emphasis on the right to be self-governing citizens. Liberal societies grant each competent adult the right to autonomy, which entails the possibility that one will make mistakes—even costly mistakes—about what makes life go well for oneself. Manipulative nudges violate that right.

Rational persuasion is an effective means to inform citizens, to engage their cognitive abilities, and to convince them that they have a self-control problem. Instead of making counterfactual assumptions about what citizens' *ideal* selves would prefer, governments should employ rational persuasion to address the limitations of *actual* selves. Such educational efforts might reveal that citizens already agree with the government on which preferences would, if fulfilled, contribute to their welfare. In other cases, citizens will change their preferences based on the arguments offered to them. A third possibility is that citizens disagree and stick to their current preferences. Attempts to bypass the deliberative capacities of this last category of citizens by means of nudging amount to a form of paternalism that is objectionable from a liberal point of view. If citizens have clearly expressed their preferences, then attempts to nudge them to act in ways *in*consistent with those preferences disrespect their status as self-governing citizens.

If, however, informed citizens acknowledge that it is difficult to act consistent with certain preferences that would contribute to their well-being, they might welcome supporting nudges. Often, it will not be very difficult to convince citizens that they, for example, eat unhealthily, that they do not save enough for their pensions, or that they should not smoke. Many times, they have long known these things. Such knowledge is not yet an all-things-considered judgment that they truly prefer to change their behaviour. They might value the taste of unhealthy food over the good effects of healthier food. Or they might not be willing to cut their budgets. However, suppose that, reflecting upon the governmental campaign, they form an all-things-considered preference for eating healthier or increasing their pension savings. They still face a self-control challenge: they have to give up enjoying the taste of bad food and must find better food. People preferring to save more have to accept a lower budget and must complete paperwork to increase their retirement savings.

Citizens with newly formed preferences but problems with self-control might welcome nudges even if choice architects are fully transparent about the underlying mechanisms. Defaults can be an example here. Employers can give employees the option to join a Save More Tomorrow plan (Thaler & Sunstein: 2009, 113–127). In such a plan, an employee's savings rate increases by a certain percentage with each pay raise. The description of the plan could involve an explanation of how the plan helps employees follow through on the intention to save more: it avoids the inertia that prevents them from taking steps to save more every time they can afford it by making the increase automatic. Instead, it exploits inertia since once in the plan, an employee will probably not make the effort to opt out unless he is truly motivated to do so; saving more in the future is easier since it involves no loss now. It seems reasonable to assume that most citizens have the ability understand how these psychological principles can work for and against their well-being and will be grateful to employ them in self-nudging strategies.

Combining rational persuasion and nudging in efforts to address social issues, then, involves distinguishing between problems of information and problems of self-control.[40] Rational persuasion provides citizens with information and arguments that provide a basis for changing their attitudes and forming their preferences. As argued above, rational persuasion does bring about argument-based attitude change. Today's attitudes towards smoking are profoundly different than they were four decades ago, and this difference has everything to do with our new knowledge about the damaging effects of smoking. However, self-control problems are a major cause of many people's continuing smoking habit, which is why self-nudging strategies are indispensable as a complement to rational persuasion. Rational persuasion is the means to take seriously citizens' deliberative capacities. Providing citizens with effective self-nudging strategies helps them act on their (re)considered preferences, and thus also takes them seriously as self-governing individuals.

This relatively strict separation between rational persuasion and subsequent nudging is crucial to combine respect for citizen autonomy with effective behaviour change. Citizens must first be convinced that they should change their behaviour. Only if they develop this insight can governments then offer nudges to help them. This procedure amounts to a minimal form of informed consent to being nudged.[41] Often this has to include an explanation of which psychological mechanisms are being marshalled in a nudge. Otherwise, citizens will not know which means of influence they are accepting. It seems possible to explain how nudges work. As humans, citizens are familiar with typical problems of self-control. Therefore, it should be possible to explain (at least to most citizens) phenomena such as inertia, loss-aversion, discounting the future, and how they can be employed in one's favour.

Some of the nudges discussed so far can complement the preceding rational persuasion that leads to informed consent, others cannot. The Save More Tomorrow plan will only become effective if employees decide to join. That decision can be the result of rational persuasion that includes an explanation of the plan's effectiveness. In contrast, as previously noted, it is effective to default employees into a retirement savings plan, even if they are not persuaded by the accompanying explanation of the plan.

Again, this is not to say that it is always wrong for governments to nudge without preceding rational persuasion. My claim is that there is a conceptual

[40] Biases related to cognitive abilities can both cause problems with proper understanding information (e.g. framing biases), and lead to self-control problems (e.g. loss aversion).

[41] For discussion of consent to nudges, see (Cohen: 2013; Holm & Ploug: 2013; Wilkinson: 2013).

difference between nudges that are complemented with preceding rational persuasion and those that are not. Of course, this difference is morally relevant since in the latter case, additional justification is needed for governments to employ them.

3.5. Conclusion

I have argued that that rational persuasion *aims* at using both arguments and non-argumentative means of persuasion so as to allow and empower individuals to change their attitudes and behaviour based on a largely unbiased reflection on arguments. From this characterization of rational persuasion, I have derived three constraints on employing non-argumentative means of persuasion. These constraints were also helpful in analysing the similarities and differences between rational persuasion and several examples of nudging, leading to more insight into the nudging approach.

Our increased psychological knowledge of human deliberation and decision-making is conducive to both successful rational persuasion and nudging. That is fortunate because a smart combination of the two often gives the best chance to solve social problems. Whereas rational persuasion may successfully lead to attitude and preference change, citizens do not always have sufficient self-control to behave accordingly. In such situations, they might be rationally persuaded to accept a nudge (or to use persuasive technology). Their status as autonomous citizens is respected while the social problem might be effectively addressed.

4 Persuasive technology redefined

4.1. Introduction

As noted in the introduction of this thesis, persuasive technologies are technologies that are explicitly designed to change human attitudes and behaviour (Fogg: 2003). And as became clear in Chapter 2, persuasive technologies raise several ethical questions, such as how they affect user freedom and autonomy, and privacy or what their unintended negative impact on the target value or other values might be. Philosophical and ethical reflection on persuasive technology, however, is still in its early stages, and a satisfactory definition or account of persuasive technology has not yet been given. In this chapter, my project is to develop a characterization of persuasive technology that builds on its current standard definition (Fogg: 2003) but remedies its shortcomings. Most importantly, the standard definition falls short in giving a clear characterization both of the distinctive mechanisms by which PTs persuade and of how PT is different from other types of technological influence on users. Of course, there is already extensive conceptual and moral reflection on influence strategies as such, but what is new is the inclusion of these influence strategies in technologies, which leads to new (moral) questions. This is primarily the case because the interaction between a persuasive technology and a human differs from interpersonal interaction in numerous morally relevant ways (as was discussed in section 2.3.1 above; see also Chapter 5 below).

I intend my redefinition of persuasive technology to be useful for ethical reflection on both existing PT and PT in the design phase. It should be acceptable to the community of researchers studying PT while allowing for criticism of their practice. In addition, my redefinition of PT should also be based on a reasonable understanding of the notion of persuasion that is faithful to the everyday use of the term and avoids overly narrow or overly inclusive definitions of persuasion.

In the next section, I will first provide a more detailed discussion of both the standard definition and its problems. In section 4.3, I will present and defend my improved account. The core of this account will be the idea that PTs persuade by communicating with the user in a way that grants the user substantial control over his mental states and behaviour. In section 4.4, I will show an important benefit of my redefinition: it enables us to make the important distinction between persuasive technologies on the one hand and manipulative, coercive, and what I will call 'limiting' technologies on the other hand.

4.2. The Standard Definition and its Problems

Following is a combination of what B.J. Fogg, the most prominent founder of the field of persuasive technology, writes about the definition of persuasive technology:

> PTs are technologies which are intentionally designed to change the behaviour, attitude or both (without using coercion or deception; persuasion implies voluntary change). (Fogg: 2003, 1,15,16).

This definition is often cited and can be regarded as the standard definition.[42]

Nevertheless, this definition has serious problems. First, it is too inclusive. For example, it counts as persuasive technology a "belief- or behavioural-disposition-inducing pill, offered and accepted voluntarily" (Nickel: 2011).[43] Obviously however, it does not make sense to describe the change of belief or behaviour caused by such pill as persuasion; doing so would clearly be alien to the common meaning and use of the term 'persuasion' (cf. Chapter 1). This over-inclusiveness originates from the second problem: the standard definition lacks a positive statement of what persuasion is. It merely states that it is not coercion and not manipulation (Nickel: 2011). However, persuasion, coercion, and

[42] This is evident from for example (W. IJsselsteijn, Kort, Midden, Eggen, & Hoven: 2006, 1) and (Oinas-Kukkonen: 2010, 6). IJsselsteijn et al. define PT as "…a class of technologies that are intentionally designed to change a person's attitude or behavior. Importantly, persuasion implies a voluntary change of behavior or attitude or both. If force (coercion) or misinformation (deception) are used, these would fall outside the realm of persuasive technology." Oinas-Kukkonen defines a behavior change support system, a type of PT, as "an information system designed to form, alter, or reinforce attitudes, behaviors or an act of complying without using deception, coercion or inducements". Like Fogg, these authors note that persuasion and manipulation cannot always easily be distinguished. However, sometimes authors who cite Fogg's standard definition omit the part that excludes coercion and deception (e.g. Narita & Kitamura: 2010, 15)

[43] This argument is made by Philip Nickel in an unpublished manuscript. I thank him for sharing his manuscript, which stimulated my reflection on this chapter's issues considerably.

manipulation are not the only types of influence; inducements[44] and incentives seem to be different types still. The procedure of defining PT by listing the types of influences it does *not* employ is too indirect. What we need instead is a characterization that positively states what type of influence PT exerts. From this positive account, distinctions from other types of technological influence will follow, excluding technologies such as the above-mentioned pill as instances of PT. A third shortcoming of the standard definition is that it can easily be misinterpreted as implying that PTs can directly change behaviour. However, behaviour change brought about by PT is always mediated by a change in some mental state. An improved definition should clarify the mechanisms underlying the possible changes of mental states.

For several reasons it is important to remedy these shortcomings of the standard definition. One is the scientific ideal of descriptive clarity, which also has practical implications (Nickel: 2011). Designers, psychologists investigating the mechanisms of technological persuasion, and ethicists reflecting on PTs sometimes have difficulties communicating with each other. This is because they sometimes have different ideas about the essence of persuasive technology and its distinctions from other technologies. Ethical reflection on PT, with the aim of guiding the responsible design of PT, depends on input from designers and other PT-scholars. Mutual understanding requires a better definition of PT.

In addition, given PT's substantial potential to help solve societal problems, it is important that citizens do not come to distrust PT in general. If, however, PT scholars use the label PT for technologies that clearly seem coercive or manipulative, they do give reasons for such distrust.[45] A clear and positive account of PT should guide the field in forgoing this inapt use of the term PT and help protect the PT 'brand'.

[44] Oinas-Kukkonen defines inducements as "exchanges of money, goods, or services for actions by the person being influenced. By definition, these are not persuasive elements" (Oinas-Kukkonen: 2010).

[45] See section 4.4 below for an example of calling PT what in fact is manipulative technology. The title of one of the contributions to *Persuasive 2012*, the yearly conference on PT, "Biometric *Monitoring* as Persuasive Technology: *Ensuring* Patients Visit Health Centers in India's Slums" (emphasis JS) correctly indicates that what we have here is not persuasion but coercion (Bhatnagar et al.: 2012).

4.3. Demarcating persuasive technology

In this section, I provide a positive account of persuasive technology that builds further on the standard definition. It does so by specifying in more detail how to understand persuasion through a technological artefact.[46] The core of this account is the idea that persuasive technologies communicate with users, as is implied by the fact that persuasion is a form of communication. In addition, the persuasive communication must be of such a nature that it leaves users sufficiently in control with regard to what they believe and what they do. For, as noted above, it is characteristic of persuasive technology that it relies on voluntary user participation. These ideas can be formulated more precisely as the following set of individually necessary and jointly (by and large) sufficient conditions:

Persuasive technologies are technologies that are (i) intentionally (ii) designed to change some mental state(s) of the user, most often with the ultimate aim of behaviour change. They do so (iii) by communicating (iv) in a way that grants users substantial control over their mental states and behaviour.

In sections 4.3.1-4.3.4, I will explain these constituent parts. At this point, it should be noted that in several cases, persuasive technologies provide task support, which functions as an additional source of persuasion. This phenomenon is so common that it merits separate treatment (4.3.5).

I should stress that the PT community's shared understanding of persuasive technology as relying on the user's voluntary behaviour change (which rules out manipulation and coercion) is the reference point of this chapter's project of redefining PT. There are good reasons to restrict persuasive technology in this way. As noted, it will help to maintain user trust in PT. In addition, if persuasive technology would be interpreted so broadly as to include technologies that manipulate or coerce users, we would lose discriminatory power. The distinctions between the different influence concepts that we have (persuasion, manipulation, coercion, etc.) would get blurred. Below, I aim to show how the

[46] I set aside the question as to how to define the "technology" part of the concept of persuasive technology. For, there is no disagreement or lack of clarity within the community of PT scholars about which kinds of technological artifacts can be persuasive technologies (e.g. websites, apps, displays, robots, etc.). For discussion of the definition of technology, see (Franssen, Lokhorst, & van de Poel: 2015; Mitcham & Schatzberg: 2009)

interpretation of PT as communicating with its users in a manner that grants them substantial control fits the basic idea of PT as relying on voluntary change.

4.3.1. Intentionality (i)

PTs are intentionally[47] designed to bring about a specific, targeted change in their users' behaviour. The notion of 'intentional design' is a key part of the standard definition that serves to distinguish PT from other technologies, a point also stressed by Fogg (2003, 16–17). All technologies influence users' behaviour, but only PTs are designed with the aim of bringing about a very specific behaviour (by means of communicating with the user, see below). In contrast, according to Tromp et al. "[p]eople who notice the influence of the microwave on their eating pattern experience persuasion" (Tromp, Hekkert, & Verbeek: 2011). However, this way of using the term 'persuasion' is atypical. The microwave was not intentionally designed to change eating patterns. It was merely designed as a tool for warming food in a quick and convenient manner. Including 'intentionality' in the definition of persuasive technology is a way to distinguish PT from technologies that unintentionally influence people in unforeseen ways. This method of singling out PTs matches the self-understanding of the field of PT research and design. (The unintentional influence of technology can be described in terms of mediation, the idea that technologies always somehow shape users' perception and action (Verbeek: 2006).)

In connection with the intentionality of technological persuasion, it will be useful to briefly mention Fogg's distinction between microsuasion and macrosuasion. Microsuasion concerns artefacts that contain designed elements of persuasion to facilitate proper use. Macrosuasion concerns artefacts that overall are designed to achieve a certain specific behaviour change, such as the BabyThinkItOver infant simulator mentioned in the introduction (Fogg: 2003). Both in this chapter and in this thesis as a whole, I will primarily be concerned with this latter type of PT.

[47] Here intentionality is used in the classical sense of a human subject that intends something, and not the in the post-modern sense according to which hybrids of human and technology can have intentions as well (for the postmodern sense, see Latour: 1992; Verbeek: 2009).

4.3.2. The target of change: mental states and behaviour (ii)

The standard definition speaks of a change of attitude, behaviour, or both. 'Attitude' in this definition should be taken in its psychological meaning: an attitude is a person's general evaluation of persons, institutions, policies, artefacts, situations, etc. This general evaluation is dependent on the person's beliefs about the attitude object. Furthermore, attitudes are important determinants of behaviour, given that they are sufficiently specific (Ajzen & Fishbein: 2000; O'Keefe: 2002).

The first thing to note is that PT can never *directly* change behaviour; as noted above, strictly speaking, the standard definition is incorrect in suggesting that it can. Because technological persuasion proceeds via communication (see below), the persuasive message must always be processed by the user and therefore will lead to a different behaviour only indirectly via altered mental states.

Secondly, attitudes are not the only mental states that determine behaviour, though, they are important ones. According to the 'Theory of Planned Behaviour' (Ajzen: 2012), a person's 'perceived behavioural control' and 'normative beliefs' (beliefs about the normative expectations of significant others, such as partners, parents, friends) also determine behaviour. For example, by giving timely feedback and helpful suggestions, an eco-feedback system can give an eco-minded driver the confidence that she can lower her fuel consumption in practice. Also, if people close to the driver express the view that one should use less gasoline, she will be motivated to do so. Another possibility is that the driver mistakenly believed that she had an eco-driving style until the moment that her eco-feedback device showed her the opposite. Together with her strong and positive attitudes towards the environment, and with the support of the eco-feedback system, this changed belief might lead to a drastic behaviour change. Therefore, in addition to attitudes, beliefs about the world, social norms, and one's capacities are mental states that co-determine behaviour and, therefore, are potential targets for intended change through PT.

Thirdly, change is a general concept that can stand for the formation, reinforcement, alteration (Oinas-Kukkonen: 2010), or activation (O'Keefe: 2002) of an attitude or other mental state. PT can be designed to aim for each type of change in each different mental state that is linked to behaviour. For example, eco-feedback systems typically do not engage in the exchange of arguments with drivers to form eco-friendly attitudes or change opposing attitudes. Instead, the relevant attitudes are activated and likely will even be reinforced as a result of the performance of eco-friendly driving.[48]

4.3.3. Persuasive technology communicates with users (iii)

According to the account of PT introduced above, communication is the positive mechanism by which persuasive technologies influence users' mental states and behaviour. This makes sense because persuasion is a form of communication, which is clear from how the term 'persuasion' is used. Whereas dictionaries vary in how they describe the meaning (s) of 'persuasion', they share the idea that persuasive influence is exerted by means of some form of communication.[49] In the previous chapter, I introduced the notion of 'means of persuasion'. In addition to arguments, a whole range of other means can be part of persuasion, e.g., emotion, appeal to authority, consensus, etc. (see also the previous chapter). Whereas philosophical analyses of persuasion focus more narrowly on the role of arguments (e.g. Benn: 1967), arguing is also a communicative process. Persuasion, then, clearly is a form of communication.

Consequently, if the term 'persuasive' is well chosen, persuasive technology should persuade by communicating with its users. If we look to the examples of persuasive technologies given in the introduction, we see that each indeed involves some form of communication. The BabyThinkItOver infant simulator generates cries to communicate her (simulated) need for care to the care-giving teenager. E-coaching websites or smartphone apps send all kinds of messages to users that suggest going for a walk, tell them what food to eat, or inform them about their friends' achievements. Most often, users interact with these sites or apps by providing a great deal of information regarding their activities, food intake, and so on. This view is confirmed by the fact that virtually all PTs described in the 2010 and 2012 proceedings of the yearly conference on Persuasive

[48] For the general idea that performing actions may strengthen congruent attitudes, see (Strack & Deutsch: 2004)

[49] The Merriam-Webster Online Dictionary describes "to persuade" as "to move by argument, entreaty, or expostulation to a belief, position, or course of action", with as second meaning "to plead with, to urge" ('to persuade': 2011). The Webster's New World Dictionary defines "to persuade" much broader as "to cause something, esp. by reasoning, urging or inducement; to induce to believe something" ('to persuade': 1988).

Technology communicate with users (Bang & Ragnemalm: 2012; Ploug et al.: 2010).

In all these examples, we see the elements of basic models of communication: a sender, message, channel, and receiver (cf., e.g., Berlo: 1960 who expands on the classic Shannon & Weaver model). Here, the PT can best be viewed as the proxy sender or the proxy agent of communication.[50] Designers and deployers have created the PT to communicate with the user independently, without continuous guidance. Therefore, PTs should not be viewed as a medium or channel. PT is not a regular communication technology through which one human can persuade another; it is not like a cell phone, a Skype connection, or the like. Instead, PTs are designed to autonomously interact with human users. In an important sense, users are 'alone with the PT'; designers or employers most often do not play an immediate role.[51] Regarding the message sent by a PT, this may be more or less elaborate and in some cases, it consists only of signals, for example, meaningful colours (such as red and green). In such a case, PT can best be viewed as engaged in 'proto-communication'. In other cases, such as carefully designed artificial agents that are able to engage in detailed and complex exchanges with human users, which may for example involve emotion recognition and artificial emotions by the PT, the interaction is much closer to full-fledged (human) communication.

The identification of communication as the mechanism by which PT changes user behaviour indeed gives us the positive account of PT that we sought. Persuasive technologies provide users with information of some sort with the aim of influencing them. For example, PTs can transfer both factual and evaluative feedback on performance, information that activates and elicits attitudes relevant to the situation, information about relevant social norms, arguments that support a certain attitude or action (e.g. Narita & Kitamura: 2010) and so on. This information is mentally processed by the user, which may lead to a mental state change; as a consequence, the user might change her behaviour.

[50] See (Nickel: 2013) for a general treatment of the idea of computers as speech actants. As long as the PT is communicating alongside the designer and deployer's script, communication is successful. If, however the PT-user interaction goes too much off-script, the communicative act has failed from the perspective of the designer and deployer.

[51] Cf. also (Fogg: 2003, 16), who stresses the difference between human-computer interaction (which is the case for PT) and computer-mediated communication.

Intuitive examples of non-persuasive technologies such as the behaviour-inducing pill, voluntarily taken, are ruled out as PT because they do not communicate. The same applies to preventing litter by making surfaces glancing, which is sometimes regarded as a form of PT (Midden: 2011). Although such surfaces change the viewer's behaviour via some mental processes, to say that the surface sends a message to the viewer not to litter is merely metaphorical speech. The glancing may be an indication or a sign that the surface is clean, but it does not have the kind of conventional meaning shared in communication.

It might be objected that the interaction between PT and the human user does not qualify as communication because only humans can communicate. Thus, it seems a mistake to identify communication as the mechanism of PT. In response, note that for the present purpose of giving a positive account of PT, there is no need to agree completely on which conditions are a necessary part of communication. I have already indicated that a PT is merely a proxy communicator. In addition, it is evident that human-human communication is much richer, subtler, and more complex than PT-human communication. Some PTs merely communicate in a highly elementary way. However, given that PTs send messages to humans in a certain context, it is meaningful to speak of PTs that communicate with humans because doing so best describes and explains how they are effective. Communication is *the* way in which PTs can persuade while granting users substantial control over their mental states and behaviour.[52]

4.3.4. Substantial control (iv)

The account defended in this chapter includes the clause that PTs communicate 'in a way that grants users substantial control over their mental states and behaviour'. This clause is needed because communication can also be used to deceive, to manipulate, or to utter coercive threats. Therefore, to distinguish PT from other technologies designed to influence users, we need to do more than singling out communication as the positive mechanism of PT. This section will clarify the concept of substantial control for the use-context of PT.[53] While

[52] Since PTs communicate with the users, the study of the ethics of communication can be made fruitful for ethical reflection on PT as well. This is done in an informative way by (Spahn: 2011), and also in Chapter 5 below. For criticism, see (Linder: 2014).

[53] Substantial control is a term familiar from the medical ethics literature (e.g. Faden & Beauchamp: 1986).

coercion and some forms of manipulation do not leave subjects much control, rational persuasion or convincing, in contrast, grants subjects the fullest possible control over their mental states and behaviour (see also the previous chapter).

Persuasive technology should be located between these extremes, and more precisely, closest to rational persuasion. In the first place, locating PT somewhere in the middle of the control continuum arguably fits best with the meaning of the term persuasion. From the dictionary entries cited above, it can be inferred that even though persuaders may derive from many different sources of influence (in addition to arguments), subjects always maintain a significant degree of choice. A well-known textbook on 'persuasion' defines the term as "a successful intentional effort at influencing another's mental state through communication in a circumstance in which the persuadee has some measure of freedom" (O'Keefe: 2002, 5). Admittedly, 'persuasion' is sometimes used in ways that are virtually the same as 'coercion' or 'manipulation'. However, these uses are unfortunate because they stretch the meaning of 'persuasion' to the extent that the concept becomes meaningless (cf. Miller: 2002). It is not without reason that we have many different terms to denote the distinct forms of influence that exist. It is better not to lose helpful and meaningful distinctions between forms of influence.

Secondly, locating PT closer to the rational persuasion end of the continuum is in line with the consensus view within the community of PT scholars. They repeatedly emphasize that PT should be designed to aim for voluntary attitude and behaviour change (Fogg: 2003; IJsselsteijn et al.: 2006; Oinas-Kukkonen: 2010). Although voluntariness itself is a much-discussed concept (e.g. Nelson et al.: 2011), 'voluntary change' at least implies that users of PT are in substantial control over their mental states and behaviour.[54]

Still, there is a significant difference between what is commonly referred to as rational persuasion and most PTs' persuasive communication. Rational persuasion is characterized by the *aim* for attitude change *on the basis of* actual reflection on arguments given (see Chapter 3 above). The influence attempt is explicit, and recipients change their attitudes (or other relevant mental states)

[54] Elsewhere I discuss this voluntariness condition on PT as the distinction mark with other influencing technologies (Smids: 2012). Here I prefer to conceptualize PT in terms of allowing users substantial control, because 'voluntary participation' has too much the connotation of an act of deliberative, conscious, purposeful user action. This connotation does not fit several PTs, for reasons that will become clear below.

only insofar as they think the reasons offered are compelling. Persuasive technologies primarily aim for behaviour change while leaving users the option to carefully reflect on what to do and not to comply. PTs are not focused on arguments but provide a much more prominent role for non-argumentative means of persuasion, such as peer recommendations, emotional appeals, social proof (Cialdini et al.: 2006), evaluative feedback on task performance (Ham & Midden: 2010), and so on, provided that they do not undermine substantial user control.

Users of the PT, then, are in a position to exercise substantial control if they can perform their *own* reflection on their reasons for action and act on the basis of that reflection. By 'reflection', I mean relatively conscious deliberation on what to do. In other words, substantial control means that PT users, if motivated, can exercise their capacities for practical rationality in such a way that they determine what is best for them and act accordingly. This explication of 'substantial control' implies three relatively global design requirements, from each of which more detailed constraints can be derived. First, PTs must *enable* the user's capacities for practical rationality. In other words, the PTs communication should support the user in determining whether the target behaviour is practically rational for her. Second, PTs should *not subvert or undermine or supplant* the user's practical rationality, for example, by providing deceptively framed or even outright false feedback. Third, the PT should also grant the user the *freedom to act on the outcome* of her deliberation.[55] Let me explain each in more detail.

First, with regard to enabling the user's capacities for deliberation, the key idea is that the PT's communication provides that which can be reasonably expected under the circumstances. Above, communication was identified as the

[55] While these constraints are developed on the basis of the previous chapter's three constraints on the use of non-argumentative means of persuasion, they do not coincide. First, persuasive technology persuasive strategies go beyond rational persuasion. Therefore, unlike rational persuasion it is not a constraint on PT to avoid persuasion via heuristic processing. Consequently, the first two constraints of the previous chapter do not fully apply. However, unless the PT communicates with the user in a way that the motivated user can make up her mind whether or not to perform the target behavior, the PT does not grant substantial control. Hence this chapter's first constraint. Put differently, whereas rational persuasion aims at argument-based attitude change, PT should at least enable such change. The second constraint on PT largely overlaps with the previous chapter's third constraint, which instructs not to "bias user reflection". This chapter's third constraint is new compared to the previous chapter, because it is a serious threat that PT disables users to act upon their reflection.

mechanism of persuasive technologies. Thus, for a PT to persuade, it must communicate in a manner that is relevant to the target behaviour.[56] This may involve giving reasons, feedback on task performance, motivational support, etc. However, these communications only support practical rationality if the PT's target behaviour and the goal(s) or end(s) this behaviour serves are known or can be easily inferred by the user. Consider, for example, the green colour of ambient lighting feedback, which could refer to a level of energy consumption below the previous week, below one's peers, below an absolute level, and so on (cf. Spahn: 2011). This target behaviour must be clear since otherwise, the green light cannot serve practical rationality.

Alternatively, consider an example that regards the aims served by the target behaviour. A health app may aim to improve the user's fitness, physical strength, and weight loss at the same time (whereas the deployer may have the underlying end of saving health care costs). Suppose that it is unclear to the user how a suggested behaviour relates to each of these aims. If he is not equally motivated with respect to each aim, he will have difficulty determining his reasons for following the suggestion. In each of these examples, the PT does not meet reasonable expectations with regard to enabling the user to make an informed decision about her behaviour. We see that the global design requirement that PTs should enable practical rationality implies relevant communicative input, along with a recognizable behavioural target and underlying aim.

The second global design requirement is also implied by the concept of 'substantial control': PTs should *not subvert or undermine* the user's practical rationality. Even if the target behaviour is clear and a PT communicates relevant information as input for deliberation, the PT might still rely on deception, manipulation, forms of peer pressure that distort deliberation, and so on. It turns out to be difficult to provide general theoretical accounts of influence types such as deception and manipulation that explain for all instances how they undermine practical rationality (Gorin: 2014; Wilkinson: 2012). Consequently, it will generally be most insightful to directly investigate the impact of specific PTs' means of persuasion on user reflection. This often requires detailed knowledge of the underlying psychological mechanisms.

[56] (Lehto & Oinas-Kukkonen: 2010) investigated six weight loss websites and found that "three out of six interventions did not reveal the purpose of the website beyond very general statements."

Nevertheless, some general guidance for such investigations can be given. A means of persuasion may undermine and subvert practical rationality by biasing, bypassing, or inhibiting reflection.[57] For example, the health app mentioned above may exaggerate the dangers of certain behaviours and undervalue the gains, which if taken seriously by the user, will bias and distort her practical judgment. To prevent this kind of biasing, designers need to ensure that feedback and other relevant information regarding the target behaviour is factually correct and sufficiently balanced.[58] Other means of persuasion that may bias and distort reflection are certain emotions (e.g., cases of fear), strong forms of peer pressure, and social consensus, to name a few.

Several of the means just mentioned may also, instead of biasing user reflection, simply exert influence in ways that bypass user reflection. As explained in the previous chapter, prominent psychological theories of persuasion distinguish between two relatively different ways of processing persuasive messages (Petty et al.: 2005; Todorov et al.: 2002). If we are motivated and capable, we will engage in effortful and slow analytic or systematic thinking about arguments and all relevant information provided by the PT. However, if we are not, our processing of the persuasive communication will be relatively effortless, quick and non-analytic, or heuristic. Often with little notice, we apply simple heuristics on cues offered by the PT. In that way, for example, the positive valence of an emotion may become associated with the target behaviour (Petty et al.: 1993). Alternatively, the stated consensus that the target behaviour is correct may lead to stronger corresponding behavioural intentions via a heuristic rule that says that "If other people believe it, then it's probably true" (O'Keefe: 2002, 150).

Of course, not all means of persuasion that bypass user reflection also undermine practical rationality. Heuristic processing sometimes leads to behaviour that, according to the user, is indeed the best thing to do. The problem, of course, is that it is difficult for designers and employers of PTs to know this in

[57] See Chapter 3 above for elaboration on these three ways and an explanation of the theoretical framework underlying this distinction. That chapter also contains detailed case studies of a few means of persuasion.

[58] Of course, our judgment regarding whether the latter is the case will differ between contexts. From persuasive health apps provided by a non-profit organization we have higher expectations than from a commercial web shop that employs persuasive strategies to sell most. But even for a web shop, some presentations of a product will just be deceptive and indefensible.

advance.[59] However, even if they could be relatively assured about what users would choose, they should not deliberately employ reflection-bypassing means of persuasion *to control* users. This intent is directly in tension with the aim of granting PT users substantial control over their mental states and behaviour. However, this controlling intent may very well be absent in the case of, for example, enhancing 'surface credibility' (Oinas-Kukkonen & Harjumaa: 2008), the analogue of persuader credibility for PT. Designers and deployers should satisfy reasonable expectations that users form in response to a credible and trust-evoking appearance of a PT. If they do so, the fact that credibility might also enhance persuasion via heuristic processing that bypasses reflection is prima facie unproblematic.

Some means of persuasion have the effect of inhibiting reflection and decreasing the likelihood that the user of a PT actively considers what to do. Positive emotions have this effect, in addition to biasing processing. The most problematic use of reflection-inhibiting means of persuasion is in combination with reflection-bypassing means. In such cases, the PT is intentionally designed to influence the user in a way that bypasses her own reflection. As a result, the user will often no longer exercise substantial control over her mental states and behaviour.

We see that substantial control implies that users know and understand which influence strategies are employed by a PT. For example, as explained in the previous chapter, ambient lighting feedback is designed to have an impact without requiring the user's focal attention. Instead, it relies on low-thought processes of which the user might be hardly aware. However, if a user knowingly uses PT that works with ambient lighting feedback and understands how that feedback is effective, she is in control and engages in what might be called 'self-manipulation'. Strictly speaking, the user is not persuaded. But because she exercises prior control over her mental states and behaviour, it makes sense to include technologies like this in the class of persuasive technologies. This is a significant choice, as it concerns a large class that the community of PT scholars views as genuine PTs (because they allow for the voluntary behaviour change taken as definitive for PTs).

[59] Instead of trying to know this in advance, designers could also adopt so-called libertarian paternalism, that aims at influencing subjects to act in ways they would act if they were fully rational. See 3.4 above, for some problems of that approach.

After this relatively extended discussion of the second global design requirement (that PTs should not undermine users' practical rationality), I will briefly discuss the third design requirement implied by the notion of 'substantial control': PTs should also grant the user the *freedom to act on the outcome* of her deliberation. It is possible that after the user decides how to act, the PT either might not allow her to act accordingly or might make doing so very difficult in one way or another. This might occur because the PT can be designed to control the interaction with the user (see 2.3.1 above and (Fogg: 2003)) such that it limits the users' behavioural options. This is especially the case if the PT provides task support on which the user is dependent in performing a desired behaviour. Imagine, for example, an online therapeutic program to reduce work-related stress, which has various built-in persuasive features. Suppose it allows the user a great deal of flexibility in goal-setting but only a very limited set of routes towards these goals that are supported by the program: it might very well be the case that the user feels too much constrained in her behaviour.

A PT could also exert an influence so strong that it is difficult to resist. For example, a social comparison function of the health app should not exert so much peer pressure that a user decides to go to the gym when she would prefer to stay home and all things considered, perhaps has good reason to stay home. Here, a solution would be to give the user the control to switch this social comparison function on and off. In any event, if the third requirement is not met, the influencing technology fails to be a genuine PT.

This discussion reveals that to ensure substantial control for PT users, designers must actively seek to meet the three global design requirements. The intent to design PTs that rely on users' 'voluntary participation' (Oinas-Kukkonen: 2010, 6) implies that designers assess how the combination of different means of persuasion exerts an influence. Moreover, they must do so for different types of users. They need to prevent situations in which, for example, communicating the majority view leads users with a less reflective nature to act in ways they would not have chosen upon more active deliberation (but again, things are different if the PT is a 'self-imposed' self-help tool).

All this does not mean, however, that designers should confine themselves to the use of arguments and factual information as means of persuasion. After all, their goal is not argument-based attitude change but behaviour change over which their users exercise substantial control. In principle, this is compatible with the use of a whole range of means of persuasion, such as suggestions, social praise, evaluative feedback, reminders, task support, etc., as long as designers actively assess how their PTs affect users. The detailed psychological knowledge they need to design effective PT will simultaneously enable them to make such impact assessments. To conclude, it is therefore reasonable to hold designers and deployers responsible for granting PT users the substantial control over their mental states and behaviour that is definitive of persuasive technology.

So far, I have discussed the four constituent parts of my account of persuasive technology: technology that is intentionally designed (4.3.1) to change the user's mental states (4.3.2) through communicating (4.3.3) in a way that grants users substantial control (4.3.4). These four conditions are individually necessary and together are largely sufficient for knowing that some piece of technology qualifies as persuasive technology.[60] I now turn to a feature that is not necessarily present in persuasive technology but is nevertheless very common and therefore merits our attention.

4.3.5. Persuasive technology can provide task support

Some PTs are designed to go beyond communicating with users in the way described above by giving users task support (see also section 2.3.1 above). In both Fogg's approach to PT (2003, Ch. 3) and Oinass-Kukkonen's concept of Behaviour Change Support Systems (Oinas-Kukkonen: 2010), task support plays a central role. The essence of the several possible forms of task support is to make the target behaviour easier to perform, making successful persuasion more likely.[61] Task support will increase both the user's actual control and, in terms of the Theory of Planned Behaviour, her perceived behavioural control.

[60] As Philip Nickel pointed out, there will be a few limiting cases which satisfy the conditions, but for which it is debatable whether they involve persuasion. An example would be a technology that merely utters simple imperatives, such as, "don't drive faster". Taken in isolation, this utterance does not for example, cite a reason in support of the target behavior, nor does it make an emotional appeal to the driver, or involve another common element of persuasive communication. However, when, for example, the technology makes the utterance with much emphasis and just at the moment the driver speeds, the driver will be reminded of moral and/or prudential reasons not to speed. Thus, viewed within its use context, this technology could be regarded as PT. I will include technologies like these in the class of PT, since this matches current practice by PT scholars, and there are no strong reasons for not including them (for example, they do not coerce or manipulate).

[61] For a general discussion about the phenomenon that technology facilitates certain actions, see (Pols: 2012).

For example, several web-shops have the option to sign up for 'one-click' buying. With one click, the product is paid for and on its way to the consumer. This adds significantly to the communicative influence strategies employed, such as recommendations based on other buyers or peers. Once visitors of the shop are motivated to buy a product, it is much easier to follow through, which enhances the chance that the product will be purchased immediately. (Indeed, we might worry whether for some impulsive persons, one-click buying makes it too easy to purchase a product). One-click donations operate similarly, although they are in the social domain instead of the commercial domain.

An important form of task support provided by PT concerns self-monitoring. For example, weight-loss websites simplify the calculation of daily calorie intake by requiring 'only' the input of food consumed throughout the day. In this way, these websites make the target behaviour, viz. achieving a healthy eating pattern, easier to perform. Of course, substantial efforts are still required, but given that one wants to monitor calorie intake, this kind of task support is crucial.[62] The so-called Quantified Self movement is unthinkable without the measurement technology that provides the data that are relevant to improving one's behaviour (Wolf: 2010).

Often, task support is an integral part of the communicative interaction between PT and user. For example, green and red ambient lights emitted by eco-feedback systems in cars persuade by communicating social approval and disapproval, respectively. In addition, they provide feedback to drivers who share the goal of driving sustainably. This is important since even if a driver is highly motivated to engage in eco-driving, she needs to know how her driving compares to optimal eco-driving. Instantaneous lighting feedback strongly enhances this learning process.

For the class of PT that provides task support, the use of reflection-bypassing and -inhibiting means of persuasion can sometimes be non-problematic. The reason is that such PTs aim at a target behaviour for which users are already motivated. This type of PT is generally used as a technological aid in reaching self-chosen aims (in which task support often plays an important role). In that case, the primary exercise of the user's practical rationality takes place before starting to use the PT. Users reflect on their aims in life and on how their

[62] In Chapter 2, I discussed how side-effects of these weight-loss website might complicate this goal of health improvement.

(habitual) behaviour is sometimes not conducive to these aims. As a result, they might want help from a PT to change their behaviour. To the extent that they understand how the PT is designed to change their behaviour, their decision to start using the PT is based on their own reflection on their reasons for action and thus qualifies as substantial control. This is the case even if the PT includes influence strategies that bypass user reflection and that otherwise would have given rise to a concern about undermining practical rationality.

4.4. Persuasive technology versus manipulative, coercive, and limiting technology[63]

Whereas all technologies affect human behaviour, persuasive technology has been identified as belonging to the class of technologies that are *intentionally designed* to influence persons. The account of persuasive technology developed above enables us to clarify its differences from other members of this class. It will become clear that each of these types of technology—manipulative, coercive, and limiting—violates some of the global design requirements discussed above that guarantee users substantial control over their mental states and behaviour.

I will be using non-moralized notions of the several types of influence. That means that the assessment of the *type* of influence is independent from and precedes its *moral evaluation*.[64] Consequently, the ethical reflection on technologies designed to influence people will involve asking at least the following two questions:

[63] This section draws from Smids (2012).

[64] Of course, persuasion, manipulation, and coercion are' thick concepts', such that using them to label a certain influence attempt already involves giving a prima facie moral judgment. However, for each of them, arguably there are circumstances in which they are and circumstances in which they are not justified. In the literature on coercion, this distinction between the type of influence and its moral evaluation is contested. See e.g. (Zimmerman: 2010). A very brief argument in favor this chapter's view runs as follows. Imagine a person S making a severe and credible threat to person A and B not to perform action *x*. Suppose both refrain from doing *x* because of the threat. Suppose further that S has a moral justification for making the threat to A, but not to B. On a moralized account of coercion, it would follow that B is coerced, but A is not, which is at least counterintuitive, since they both feel forced to refrain from doing x because of the threat.

i) What is the influence type employed by the technological artefact? (conceptual or meta-ethical)
ii) When and how can this use be morally justified, if ever? (applied ethical)

This chapter is only concerned with the first question, which precedes the second. Of course, the questions are related. Whereas in principle, persuasion (and persuasive technology) is a method of influence that respects user freedom and autonomy, this is not the case for manipulation and coercion. Since the latter both infringe on freedom and autonomy, explicit and sufficient justification is required. Thus, it is clearly informative for the ethical evaluation of influencing technology to determine whether it is persuasive, manipulative, coercive, or limiting (see below).

Manipulative technology

As noted above, it is difficult to provide general theoretical accounts of manipulation. Typically, manipulators attempt to control the manipulatee by subverting her reflection in a manner that is difficult to detect. Manipulation therefore implies that the second global design guideline is violated. For example, there is technology that enables what the authors call 'unconscious persuasion' (Ruijten, Midden, & Ham: 2011).[65] The authors show that research participants who are primed with the goal 'to perform well' performed significantly better on an experimental task under the influence of subliminal feedback compared to participants in the condition of no feedback. The feedback consisted of a 17-millisecond flash of either a sad or a happy face, dependent on whether participants made the wrong choice or the right choice.

Importantly, in this study, the goal was primed, i.e., it was created in such way that the participants even were not aware of having this goal. Of course, this need not be the case; subliminal feedback on consciously self-chosen goals is also possible. In any event, users in the experiment could not be aware of the subliminal stimulation and were influenced in a manner they could not control.

We could modify the case by supposing that the technology also provides arguments for the target behaviour upon which the users can reflect. Even then, it is not entirely for them to decide how to act on this reflection because subliminal feedback is a partial cause of action. Again, the second design requirement is

[65] For subliminal 'persuasion' see also (Barral et al.: 2014; Dijksterhuis, Aarts, & Smith: 2006).

violated, and subliminal feedback cannot be part of persuasive technologies, but instead is a form of technological manipulation.

Coercive technology

Coercion consists of *A* making a credible threat to inflict some serious harm on *B* if *B* performs (or refrain from performing) action *x*, where *B* chooses to act to avoid the harm. Thus, B does not act how he would have acted if there had been no coercive threat. In that case, he would have considered his own reasons to perform x, reasons that relate to *x*. The threat provides him with an external reason, the fact that doing *x* would cause *A* to seriously harm him, which does not bear on the question of whether or not *x* itself is a good thing to do (cf. S. Anderson: 2011).

Following this analysis, seat-belt reminders in cars that continue to blink and beep unless you fasten your seat-belt are coercive. The lights, often showing a seat-belt symbol, remind you to fasten your seat-belt and may cause you to reflect on your reasons for doing so (insofar as you are familiar with them): your own safety, the safety of others, and the prevention of an economic loss to society. If these considerations are not enough to motivate you to fasten your seat-belt, likely you will fasten it nonetheless because you cannot bear the irritating sound. At minimum, you regard the inconvenience of the sound as greater than the inconvenience of fastening the seat-belt. Even though seat-belt reminders communicate with users, they violate the third global design requirement. As a driver you are not free to act on your own reflection of whether you want to wear a seat-belt, but you are provided with a decisive incentive to do so, namely, irritating sounds (and lights). Coercion cannot be part of a genuine persuasive technology.

Another example of a coercive technology is that of speed bumps. Drivers know that if they drive over them too fast, they will damage their cars. This threat is enough to slow down even the most stubborn speeders, who certainly would not have slowed down for a mere traffic sign indicating the maximum speed. These drivers are, therefore, coerced.

Limiting technology

Limiting technology is technology that makes certain behaviour impossible. This is done by reducing the number of options or even limiting them to the only one that is desirable or acceptable (according to the designer and/or deployer). For example, the controlling variant of the Intelligent Speed Assistant for cars makes it technologically impossible to exceed the speed limit. Based on localizing technology and a data system that contains all local speed limits, the technology withholds the option to speed.[66] The Intelligent Speed Assistant will also communicate speed limits to the driver, but this information is redundant in light of the technological exclusion of the option to exceed the limit. Unlike persuasive technology, limiting technology is not designed to bring about desired behaviour by communicating reasons for certain behaviour; it simply makes it technically impossible to behave otherwise. Therefore, it is clear that the driver has no freedom to act on his own reflection about his reasons for action and here as well, the third global design requirement is violated.

It is worth noting the difference from coercive technology. For example, in the case of the speed bump, the option to drive too fast is not excluded by technological means. Anyone who is prepared to damage his car can do so, while Intelligent Speed Adaptation makes it physically impossible to drive too fast. Whereas coercion operates via users' psychological states, limiting technology does not.

This discussion makes it clear that coercive, manipulative, and limiting technology—each in a different way—prevent users from acting on their own deliberation about what to do. Even if some communication is involved, these types of influences do not make themselves dependent on the outcome of the user's own choice and voluntary cooperation. Instead, each type of technology involves its own mechanism to ensure (to the greatest extent possible) that the target behaviour is performed. Therefore, persuasive technology is fundamentally different from these other types of influencing technologies, even if in practice it may sometimes be difficult to determine the class to which a specific technology belongs.

4.5. Conclusion

This chapter addressed the problems of the standard definition of persuasive technology, most importantly its failure to identify the mechanism of persuasive technology. I provided a redefinition of persuasive technology and argued for and explicated the following account:

[66] In cases where limiting technology is employed to bring about compliance to the law, it is often called techno-regulation (Brownsword: 2005) see also Chapter 6 for an extensive discussion of Intelligent Speed Adaptation.

Persuasive technologies are technologies that are (i) intentionally (ii) designed to change some mental state(s) of the user, most often with the ultimate aim of behaviour change. They do so (iii) by communicating (iv) in a way that grants users substantial control over their mental states and behaviour.

Task support was discussed because it is a frequently used additional source of persuasion, one that in principle satisfies the condition of substantial control. I developed three global design requirements that should guarantee that users of a PT indeed remain in substantial control of their mental states and behaviour. First, the PTs' communication should *provide information that supports* the user in determining whether the target behaviour is practically rational for her. Second, PTs should *not subvert or undermine* the user's practical rationality. Third, the PT should also grant the user the *freedom to act on the outcome* of her deliberation.

This account satisfies the adequacy conditions specified in the introduction. The mechanism of persuasion was identified by noting that communication is the central element in our everyday notion of persuasion. Persuasive technologies likewise persuade by communicating to users, which may include providing arguments, factual feedback, peer recommendations, social norms, etc. My account, then, is based on a plausible understanding of persuasion.

This should also contribute to rendering it acceptable to the community of scholars working on PT. In my account, the vast majority of technologies regarded as PTs are indeed classified as PTs. However, the three global design requirements, which should guarantee users substantial control over their mental states and behaviours, sometimes rule out technologies. Thus, my account also enables criticism of the practice of PT scholars.

Moreover, my account supports improvement of the development process of PTs. Importantly, my account explicitly assigns designers a relatively active role and expertise in ensuring that the PT they design indeed meets the requirements. They should use their knowledge of psychological mechanisms of persuasion that they inevitably need to design effective PTs to ensure that users retain substantial control. In this way, ethical considerations can have the largest impact because they are incorporated into the design phase and thus make a difference for the resulting persuasive technologies (cf. Friedman & Kahn: 2003; van de Poel: 2009).

In the next chapter, I will discuss a class of persuasive technology that will illustrate the constituent parts of this chapter's improved definition of persuasive technology. More specifically, the next chapter will provide a very detailed discussion of one non-argumentative means of persuasion, namely, the use of similarity between persuasive technology and the user to induce trust and enhance persuasion.

5 Buying more from an agent that looks like you. The ethics of similarity in persuasive technology.

5.1. Introduction

Whenever humans interact, they exert social influence on each other. For example, we automatically mimic each other's posture, speech patterns, and facial expressions, and usually are unaware of it. Mimicry is an example of 'similarity', which is one of the several forms or sources of social influence between humans.[67] Here, similarity can refer to any way in which two humans may be similar to one another: body posture, dress, personality, values, interests, ethnicity, gender, face, etc.

Similarity as a source of social influence can also be incorporated into the design of persuasive technologies to increase their persuasiveness (Ham & Spahn: 2015; Verberne: 2015). These persuasive technologies could be artificial social agents, which are computer programs that can interact with human users while displaying elements of 'social behaviour', e.g., giving praise, thanking, turn-taking, etc. One step further go so-called Embodied Conversational Agents, which are "computer-generated cartoonlike characters" able to engage in human-like face-to-face conversation (Cassell, Sullivan, Prevost, & Churchill: 2000, p. summ.).[68] In principle, these persuasive technologies could be designed to be

[67] An well-known classification is the one by Robert Cialdini (2006), It distinguishes 'reciprocation', 'commitment and consistency', 'social proof', 'liking' (which includes 'similarity' as an important source or cause of 'liking'), 'authority, and 'scarcity'. Other classifications are possible, depending on the level at which the forms are identified, e.g. the phenomenal level (which seems to be the case for Cialdini's classification), or the level of explanatory mechanisms.

[68] In the literature, the term 'avatar' is also often been used. However, strictly speaking an avatar is not an artificial agent, but instead 'being operated' by a real human, who is, so to say, hiding behind his avatar. In this chapter, I will use terms like "artificial agent" and "digital agent" to refer to both artificial social agents and to embodied conversational agents.

similar to human users in any possible way that humans can be similar to one another, given technological feasibility.

However, this immediately raises a concern about manipulation. The psychological mechanisms by which similarity enhances persuasiveness between humans largely operate at the unconscious level. The intentional use of such mechanisms in the design of persuasive technology with the aim to persuade human users clearly seems to be manipulation: they are influenced in ways that bypass their own reflection and decision-making (see 4.3.4 above). Consequently, such persuasive technology does not grant users substantial control over their mental states and behaviour, and thus is not genuine persuasive technology.

In this chapter, I will investigate how we should evaluate the use of similarity in persuasive technology from an ethical point of view. Is it possible for designers and deployers to make responsible use of similarity and if so, how? This ethical reflection is urgent since research in social psychology and human-computer interaction reveals a growing potential for enhancing the persuasiveness of PT, which has many potential applications in, e.g., e-commerce, education, and health care (Bickmore, Schulman, & Sidner: 2013; Holzwarth, Janiszewski, & Neumann: 2006; McGoldrick, Keeling, & Beatty: 2008; Roubroeks: 2014; Wang, Baker, Wagner, & Wakefield: 2007). More specifically, there is an increasing literature on the influence of digital agents that display similarity. People have more trust in a character whose face is digitally morphed to resemble their own face (Bailenson, et al.: 2006; DeBruine: 2002), regard digital agents that mimic their head movements as more persuasive (Bailenson & Yee: 2005), and perceive a chat robot that mimics their response time as more intelligent (Kaptein, et al.: 2011).

This chapter's main argument will be that to be morally acceptable, the use of similarity in PT must comply with the design guidelines developed in the chapter. If these guidelines are not met, the result will often be a piece of manipulative technology. In section 5.2, I will discuss the psychological mechanisms that underlie the influence of similarity. This will clarify that the nature of the psychological mechanisms that underlie the influence of similarity on us gives rise to the concern about manipulation. To further and more specifically investigate this worry, I will apply elements of Habermas's ethics of communication (section 5.3). On that basis, I will argue that similarity-induced influence is first and foremost morally problematic if its use raises reasonable and justified expectations in users, which are nevertheless unmet by how the persuasive technology is designed. In section 5.4, I develop three guidelines that should result in the morally responsible incorporation of similarity in persuasive technology. These guidelines are so stringent that designers and deployers must be very careful if they wish to use similarity. In section 5, I discuss several objections that might be raised against this chapter's overall argument. I conclude that although manipulative artificial agents are a real danger, the responsible use of similarity is possible in the design of many socially valuable applications of persuasive agents.

5.2. Similarity and 'computers as social actors'

We can begin to gain a better understanding of the use of similarity in persuasive technology by noting that humans display a wide variety of fundamentally social responses to computers. Many of these social responses were shown in studies by Reeves, Moon, Nass, and colleagues, a number of which are discussed in (Fogg: 2003; Nass & Moon: 2000). For example, participants clearly applied gender stereotypes to computers with voice output. Both male and female participants rated a male-voiced computer as more informative than a female-voiced computer about a topic categorized as masculine (computers) and female-voiced computers as more informative about a 'feminine topic' (love and relationships), even though the same information was presented in both cases. Also, people were more polite to a computer when it asked for evaluations of itself than they were when that computer asked for evaluations of another computer. In short, humans treat computers as social actors.

Nass and Moon mention three 'cues' that may lead humans to perceive computers as social actors: words for output, interactivity, and human-role fulfilment. Importantly, they argue that people respond *mindlessly* (non-reflectively and automatically) to these social cues. They "...apply social scripts – scripts for human-human interaction – that are inappropriate for human-computer interaction, essentially ignoring the cues that reveal the essential asocial nature of a computer" (Nass & Moon: 2000, 83–84). Participants in the studies mentioned above were unaware that they treated computers as if they had a personality or as if they were a member group (where they merely wore a blue armband, the computer had a blue border around its monitor, and they were told to be in the blue team with the computer).

Given that humans treat computers as social actors, it comes as no surprise that computers and hence persuasive technologies can gain influence by appear-ing similar in some way to the user. To understand a PT's similarity-based influence on humans, we first need to obtain an understanding of the role of similarity in human-human interaction. Since the psychological mechanisms underlying different types of similarity may differ, I will discuss mimicry as a case study to gain a deeper understanding. Mimicry is an important type of similarity and the impact of mimicry on PT-human interaction is relatively well studied.

Mimicry can be characterized as nothing more than "copying another's observables" (Chartrand, Maddux, & Lakin: 2006, 335). Mimicry between humans can take several forms: we may mimic each other's posture, gestures, speech pattern, accent, mood, head movements, and the like (Cf. also Maddux, Mullen, & Galinsky: 2008). This matching of each other's behaviours generally occurs unintentionally and goes unnoticed by the interacting humans. "We simply seem to have an innate tendency to do what others do" (Maddux et al.: 2008, 462). One explanation for this important feature of mimicry can be found in the 'perception-behaviour link', which refers to the unintentional and unconscious effects of social perception on social behaviour: people automatically behave as they perceive (Chartrand et al.: 2006). Applied to mimicry, this means that perceiving the observable aspects of others (expressions, postures, behaviours) activates the associated representations in memory, which makes the same action more likely.[69]

Mimicry is found to have several effects that indicate a common or general function of creating and strengthening social bonds. Chartrand et al. (2006, 348) speak of mimicry as "social glue" that facilitates our need to get along with each other. More specifically, mimicry leads to enhanced feelings of affiliation, enhanced liking, and better cooperation in some negotiation settings. Furthermore, people tend to show more pro-social behaviour after being mimicked. Conversely, having prosocial attitudes increases mimicry, which shows that there is a bidirectional relationship between mimicry and prosocial attitudes (Chartrand et al.: 2006; Leighton, Bird, Orsini, & Heyes: 2010; Maddux et al.: 2008).

[69] Social psychology also gives a background theory here: so-called interpretive for perceiving and interpreting actions overlap with behavioral schema's for producing actions and therefore perception leads to action, whereas action leads to different interpretation (Chartrand, Maddux, & Lakin: 2006, 346–347). Cf. also (Strack & Deutsch: 2004)

Notwithstanding the automatic character of typical mimicry (as explained by the perception-behaviour link), the extent to which people mimic depends on more than just perceiving another's behaviour. In other words, the extent of mimicry depends on human needs and goals. If people have a need to affiliate, they unconsciously mimic others to increase liking and affiliation. Conversely, psychological evidence suggests that disliking others decreases our mimicking of them. The point is that a need such as to affiliate need not be consciously pursued, but it may automatically trigger mimicking behaviour. This is a form of what has been called 'automatic goal pursuit', a type of automatic self-regulation that we need because of our limited cognitive capacities (Bargh & Chartrand: 1999).

Whereas people most often mimic each other without being aware of it (even if they have the goal of affiliating), we can also *intentionality* mimic other persons to achieve strategic goals. In psychological experiments, waiters received larger tips when they verbally repeated clients' orders, as they were instructed to do (van Baaren, Holland, Steenaert, & van Knippenberg: 2003). Negotiators who deliberately mimicked their negotiation partner were able to create more common value and to gain a greater part of that value. Interestingly, it was found that "interpersonal trust mediated the relationship between mimicry and deal making". Thus, the mimicking negotiation partner was trusted more, which led to better cooperation and better outcomes, with the mimicking partner receiving the greatest benefit (Maddux et al.: 2008). Finally, subjects who were given the task of mimicking other subjects were better liked (Chartrand & Bargh: 1999).

From the perspective of the paradigm of computers as social actors, it was to be expected that research indeed shows that like real humans, *digital* agents who mimic people also gain social influence. Participants in experiments regarded digital agents that mimicked their head movements as more persuasive (Bailenson & Yee: 2005), although other studies report less unequivocal results (Hale & Hamilton: 2016; Verberne, Ham, Ponnada, & Midden: 2013). Also, participants perceived a chat robot that mimicked their response time as more intelligent, whereas perceived intelligence is hypothesized to enhance persuasion (Kaptein et al.: 2011). It seems that the same perception-behaviour link that operates between humans is also activated in humans when they perceive artificial social actors that mimic them. This conjecture matches well with Nass and Moon's thesis that people respond mindlessly, that is, unconsciously and automatically, to the same cues, whether exhibited by humans or by computers.[70]

In summary, the present discussion of mimicry, serving as an example of similarity, illustrates several things. First, mimicry is a very natural form of behaviour that usually occurs automatically and unconsciously, although it serves the general purpose or function of what we might call 'social cooperation' by enhancing liking, trust, affiliation, and the like. Furthermore, mimicry can also be employed more consciously and deliberately such that it delivers strategic advantages in social cooperation. Finally, analogously to this latter way of employing mimicry, mimicking artificial social agents also seem to be more influential on humans. Thus, it seems that the purpose or function of mimicry is preserved when mimicry is incorporated into persuasive technology. As noted, the psychological mechanisms underlying other forms of similarity influence between humans and between a PT and a human need not be the same as they are for mimicry. However, in line with general findings in social psychology and the 'computers as social actors' paradigm, these mechanisms will often have an automatic, unconscious and unreflective nature. This feature renders ethical reflection desirable, as will be discussed in the next section.

5.3. Similarity, manipulation, and an ethics of communication

The previous section gives ground to the concern about manipulation in cases in which similarity is incorporated into PT to enhance its persuasiveness. The psychological processes through which similarity has an influence between humans are mainly automatic and they bypass our reflection. Because designers and deployers intentionally use these means to influence, they seem to manipulate users of PT. Since manipulation can be broadly defined as the type of influence in which the manipulator intentionally manoeuvres someone by means of either bypassing or distorting her reflection and decision-making, the normal exercise of one's capacities for self-government experiences interference (Baron: 2003; Cave: 2007; Gorin: 2014; Nettel & Roque: 2011; Noggle: 1996; Sher: 2011; Wilkinson: 2013).

[70] These are my own observations. The studies cited are silent about the underlying mechanisms in case of the PT-human interaction and merely refer to the literature on mimicry between humans.

According to most ethical theories, manipulation is prima facie wrong, absent special justification. Deontological or Kantian approaches will most directly condemn manipulation as a breach of person's right to autonomy, to self-government. Consequentialist approaches will point to negative consequences of manipulation, such as distrust, less cooperation, anger, and so on. Virtue ethical approaches explain which vices are at stake in manipulation (Baron: 2003). Therefore, it seems as though the use of similarity in PT can straightforwardly be regarded as morally wrong because it amounts to manipulation.

However, the charge of manipulation may be raised too quickly, or in a too general way. Suppose designers merely aim to build an agent that is capable of the same amount and type of social influence based on similarity as humans are. In that case, things that humans do *un*intentionally and automatically as a result of their psychological make-up, designers *necessarily* must incorporate in their agent *intentionally*. If, for example, I am influenced in exactly the same way by a fellow human and a digital agent mimicking me, the second but not the first case would be manipulation. The fellow human's use of mimicry is unintentional, whereas the agent's use of mimicry is the result of designer or deployer intentions. However, since it seems at least somewhat counterintuitive to think of digital agents as manipulating us if they do just what humans do, we must engage in further reflection.

For this purpose of developing a more fine-grained analysis of whether and when the use of similarity in PT is manipulative, I will employ notions from both Habermas's speech-act theory and his ethics of communication (ccf. Spahn: 2011). According to discourse ethics, communication is inherently normative. Whenever a person performs a speech act, she raises three implicit validity claims (Habermas: 1985; Habermas: 1998; McCarthy: 1981):

i) that the utterance is true,
ii) that the utterance truthfully expresses the intentions or the 'inner nature' of the speaker so that the hearer can trust the speaker (also called the 'sincerity condition'), and
iii) that the utterance is appropriate with respect to prevailing norms and values.

In earlier work, Habermas also included 'comprehensibility' as a validity claim, but now he holds that comprehensibility is a condition on an utterance being communicative.[71] Because the utterances of PTs should be comprehensible as well, I will include this condition in the discussion.

Both the listener and the speaker are aware of these validity claims as necessary background conditions for successful communication. If one of the claims is not met, the result is failed communication of some sort, be it misunderstanding, deception, or manipulation.[72]

I will first argue that it makes sense to apply the notion of raising validity claims to a PT's and particularly to a persuasive agent's communication with users. Strictly speaking, agents cannot raise validity claims themselves.[73] To argue that this indeed makes sense, I will first make it plausible that agents can be designed in a way such that they meet all three validity claims. Second, although agents cannot raise validity claims themselves, users will nonetheless respond to agents as if they do so, and are justified in doing so.[74] To start with the comprehensibility condition, if an agent's utterances are incomprehensible, human users clearly will not understand what the agent attempts to communicate. The existence of agents that successfully exchange text messages with

[71] Thanks to Joel Anderson and Andres Spahn for help on this point.

[72] Strictly speaking, Habermas gives an analysis of what he calls communicative action, in which speaker and hearer reach mutual understanding. Communicative action takes place in the 'lifeworld', the world of everyday life in family and society, and not so much in other domains, like markets, in which communication is more of a strategic nature (Bohman & Rehg: 2014). In strategic action, both interactants know that the aim is not the reaching of mutual understanding, but achieving individual success. Thus, the application of an Habermasian ethics of communication to agents may be more or less fitting for different use contexts of these agents. But still, every successful speech-act is bound to meet the validity claims because a speaker who does not meet them makes the 'performative contradiction'. That is, she is contradicting the very conditions that must obtain for the performing of the speech-act to be possible. For example, if I make a promise while having the intention not to keep it, I am insincere whereas the institution of promising can only exist under the condition of sincerely expressed intentions. Austin qualified such speech acts as abuses, one of the two ways in which speech acts may fail to be felicitous (Green: 2009). Thanks to Andreas Spahn for helpful discussion.

[73] Is it in fact the designer or deployer communicating with the human user with the agent merely as a medium? It seems to be more accurate to say that the agent is designed to communicate with the user, even if the communication will never be as rich and complex as between humans. Once the agents interacts independently with users, the designer does nothing. So, the agent is not a medium in the same way as a telephone or skype, or an avatar. See also Chapter 4 above.

[74] See Spahn 2012 and Chapter 4 above for a general defense of the idea that persuasive technology can establish a communicative relation with its users.

humans proves that designers are capable of ensuring that agents utter intelligible sentences. In addition, comprehensible agent utterances are true or false with regard to their propositional content or existential implications of that content (claim to truth). With regard to the validity claim of truthfulness, even though agents do not have intentions themselves, their utterances will express information and cues about how these agents are most likely to deal with their human users. If, for example, a medical digital agent explains to the patient that the answers to the questions he is going to ask will be answered confidentially, the patient will justifiably assume that the agent is designed to fulfil that promise. Alternatively, if the human user is asked to communicate some of her preferences related to a certain type of product that she is considering buying, then a truthful agent is designed to take these preferences into account in the subsequent interaction in a way that takes them seriously. As we see, utterances can vary in the degree to which they truthfully represent how agents are designed both to interact with humans and to treat them in specific cases. Finally, agents should be designed to make utterances that are correct or appropriate with respect to accepted norms and values (claim to appropriateness). The medical digital agent that makes a joke related to a patient's serious illness will be distrusted by many users, who may refuse further interaction.

From the discussion so far, it is apparent that human users generally adopt an attitude towards agents that assumes that they indeed meet the three validity claims. Generally, users will expect agent utterances to be true and to give a truthful indication of what agent 'behaviour' they can expect, and they may be even more surprised by a socially inappropriate agent utterance than by its human analogue. To a substantial extent, this will be an unavoidable consequence of our psychological make-up: we just happen to respond to speaking artificial agents in that way, usually without much reflection. Only if we have good reason to believe that an agent will not meet the validity claims, e.g., because it is the agent of a highly untrustworthy institution, will we deliberately adopt a different attitude and potentially choose not to interact with the agent.

However, going beyond mere prediction, users are *entitled* to satisfaction of their validity claims by communicating digital agents; they *legitimately expect* this from such agents and from the designers, who will see to it that this is indeed the case. Designers and deployers know that users will take this stance towards agents. They even actively attempt to bring it about since the success of their digital agents depends on human users' interaction and cooperation. Because designers aim to bring users to assume that their artificial agents satisfy the validity claims and profit from this assumption (which leads users to trust the agents), they have a duty to make this assumption justified and rational.[75] In other words, they have to design their agents such that their utterances are indeed comprehensible, true, truthful, and socially appropriate. If designers fail to do so, they are responsible for building persuasive artificial agents that, in Austin's terminology, perform infelicitous speech acts: i.e., speech acts that misfire because they are not taken up by users or even speech acts that are abusive of the users if they, for example, generate trust without acting correspondingly (cf. Green: 2009). Insofar as designers aim to steer users towards behaviour from which they as designers profit, but their users do not, they are engaged in outright manipulation. We might say that this type of manipulation is parasitic on the communicative success of digital agents.

I will now turn from applying Habermas's validity claims to communicating agents in general to the use of similarity in persuasive agents or persuasive technology more specifically. It requires some extension and interpretation to make Habermas's ethics of communication applicable. Habermas makes clear that in his analysis of communicative action, he ignores "nonverbal actions and bodily expressions", although he agrees that these are part of communicative action (Habermas: 1998, 21, 62). However, we can still see whether it makes sense to apply the validity claims to certain types of similarity which involve such nonverbal communication. Take, for example, a person who intentionally mimics another to induce that person to like him and cooperate with him, whereas that second person has no such aims himself whatsoever. Indeed, the mimicking person wants to achieve certain strategic aims by using mimicry that benefits him, but not the other person. This person's nonverbal communication clearly does not truthfully represent his intentions since, as explained in the previous section, mimicry normally originates either from the aim of affiliating or from liking the other person,[76] or simply as an automatic response to another person's mimicry. Therefore, it is plausible to interpret such nonverbal com-

[75] For a discussion of trust as a normative expectation and a moral attitude (ascribing moral obligations to the trustee to perform in accordance with the trust placed in them) see the work of Philip Nickel. Nickel also argues for the idea that users can trust technology and be *entitled* to its performance in a way that renders appropriate their reactive attitudes of anger and blame in cases the technology fails (Nickel: 2011; Nickel: 2009; Nickel, Franssen, & Kroes: 2010). We could say that the idea of trust in technology becomes all the more plausible for artificial social agents which are designed to play many of the roles of real humans.

[76] But remember that we are usually not aware of our mimicry arising from these aims.

munication as a violation of the second validity claim: the utterance, including its nonverbal parts, does not truthfully represent the speaker's intentions. This behaviour amounts to manipulation since the speaker deceives the other person about his "own strategic attitude" (Habermas: 1998, 62, 93).

What is one to think of the analogous case for a persuasive digital agent? Above I interpreted the claim to truthfulness of agent utterances as follows: utterances should truthfully represent how the agent is designed to interact with humans and to treat them in specific cases to which these utterances are relevant. Thus, whereas in human-human communication, truthfulness means that the communication reveals the true intentions of the speaker (such that the hearer can trust the speaker), in artificial agent-human communication truthfulness can be taken to mean that the agent is designed to treat the user in a way that the user reasonably infers from the agent's communication (recall the examples given above, e.g., promising confidentiality with regard to medical information).

Analogous to the human mimicker, a persuasive agent that mimics users' head movements merely to be more influential, but without being designed in any way to be cooperative or beneficial to the user (in whatever way that can be specified), does not truthfully represent its design; its true design and the true intentions of its designer are hidden. Thus, it equally violates the validity claim to truthfulness implicitly raised by the agent's communication and taken up by the user accordingly. The experiment with computers presented as the teammate mentioned above (section 5.2) would also count as a violation of the truthfulness conditions. The computer's blue colour was meant to (nonverbally) communicate membership of the same team. Users were induced either to believe or to tacitly assume that the computer acted as a teammate, but in reality, nothing in the design corresponded to the computer actually being a teammate (regardless of what exactly that could mean). Thus, where human users perceive similarity in the sense of being members of the same team, they are deceived and to the extent that they respond more cooperatively to the computer, they are manipulated.

To conclude, the use of similarity to make PT more influential on users can be assessed using Habermas's ethics of communication. In particular, certain uses of similarity will contradict the validity claim to truthfulness upon which every successful speech acts depends. If the agent is not designed to interact with users in the same manner as users tacitly infer from the similarity used or are more or less unreflectively induced to expect, then its communication is untruthful.

As a result of looking at the use of similarity in PT from the perspective of an ethics of communication, we have acquired a more nuanced take on the manipulation charge. Only uses of similarity to design for persuasiveness are straightforwardly manipulative, thus violating the validity claim to truthfulness. However, as long as the role of similarity in the agent's communicative interaction with human users truthfully represents how it is designed to interact with its users, the charge of manipulation may be dispelled. In the next section, I will develop this insight for purposes of formulating the first of three design guidelines for the responsible use of similarity influence in PT.

5.4. Design guidelines for the use of similarity in PT

Having treated this chapter's first question about how to evaluate the use of similarity in persuasive technology, in this section I turn to the second question: how exactly can designers ensure that their use of similarity in PT does not manipulate users and is morally responsible?

In this section, I will develop three design guidelines that *together* should result in the responsible use of similarity in PT. After introducing and discussing the guidelines, I will briefly explain how they are jointly successful in their task.

The first guideline follows naturally from the idea that the use of similarity should not make the persuasive technology, often artificial agents, untruthful in their communication with users. Noting that persuasive artificial agents are designed to interact with humans and to treat them in specific ways, we can formulate the first guideline:

Design guideline 1: Similarity should truthfully represent or indicate how artificial agents are designed to interact with users (abbreviated as 'G1 truthful design').

Again, the underlying idea is that users will tacitly infer beliefs about how agents will deal with them based on certain ways in which the agent conveys similarity in its communication with those users. In some cases, even the phrase 'tacitly infer' implies too much reflective awareness of similarity. Some forms of similarity will more directly *cause* certain user attitudes, such as trust and willingness to cooperate, by way of mainly automatic cognitive processes. In the terminology of section 5.2, users will respond “mindlessly”. ‘G1 truthful design’ instructs designers not to exploit user attitudes such as trust and cooperativeness, but to make it rational or justified for users to have them. The only way to do so is to design the agents to be actually trustworthy, which would make it worthwhile for the user to be cooperative. In this manner, designers and deployers ensure that the design meets these user beliefs and expectations they infer from agent similarity.

‘G1 truthful design’ rests on two assumptions that must be made explicit. The first assumption is that it is clear or can be determined how the several types or sources of similarity are interpreted by users or how they more directly cause certain user attitudes. These interpretations or attitudes should be highly uniform across users so that it can be possible for designers to adapt the design to all users (but see section 5.5 below). The second assumption is that designers are indeed able to adapt their design to user expectations and attitudes resulting from the use of similarity.

To illustrate ‘G1 truthful design’ and confer some plausibility on these two assumptions, consider the following example of a sales agent in a web-shop for mountaineering equipment. This agent is designed to assist customers and sell them products; it is a piece of PT. In a physical store, it is not unusual for a salesperson to wear mountaineering clothes, and nothing seems wrong about doing so.[77] With this intentional and deliberate choice of clothes, the salesperson may have several aims: to recommend the brand of the clothes, to suggest her expertise, and to suggest belonging to the same group of mountaineering fans as the client. The last aim relates to what psychologists refer to as in-group favouritism, which seems to operate largely via automatic psychological processes of which people are unaware. Generally, membership in the same group leads to more trust and cooperation (Hogg, Hohmann, & Rivera: 2009; Tajfel, Billig, Bundy, & Flament: 1971).

Is this salesperson manipulating clients? Not insofar as her wearing the clothes is a natural thing for her to do as a sincere mountaineering fan and truthfully represents her intentions. Even if she deliberately aims to influence clients by showing herself as ‘one of them’, she also provides clients with good reasons to trust her and her advice. As a mountaineer, she knows the important features of good gear, she is better able to understand clients and she can give

[77] Thanks to Philip Nickel for the example.

better advice. Thus, although the in-group influence may impact consumers directly, without considering the reasons just noted, these reasons justify this impact, at least to some extent. If, however, the salesperson knows very little about the sport and would not like it at all, but for commercial reasons aims to give clients the impression that she is a fan, then wearing mountaineering clothes would be an untruthful representation of her intentions, and thus she would be manipulating.

What should we think of the digital equivalent, a sales agent for a web-shop that is 'dressed up' as a mountaineer? Suppose that the designers have the same aims as the human salesperson: to recommend the brand of the clothes, to suggest expertise, and to suggest being a member of the same group of mountaineering fans as the client. To influence consumers, designers deliberately raise consumer expectations that the sales agent is an expert, likes mountaineering, and knows what is important in the sport; thus, agents are bound to meet these expectations. Of course, this applies only to the behavioural level of interaction with the users. The artificial agent is not a human and therefore cannot like mountaineering, but it can behave in ways that are fitting for someone who does. Insofar as designers can meet these self-created expectations, their agent's communication truthfully represents how it is designed to deal with users.

Note that in this example, the assumptions underlying 'G1 truthful design' appear reasonable. It seems possible for designers to know how users interpret the agent's dress; indeed, that knowledge is the very reason to dress the agent as a mountaineer. Also, it seems perfectly possible to design an agent who is indeed knowledgeable and adapts to the customers. Some even argue that e-commerce agents can better adapt to customers than human salespeople in several respects (McGoldrick et al.: 2008, 451). What should designers do in cases where there is good reason to doubt that they know how users interpret similarity or which attitudes users are induced to adopt through similarity? In such situations they cannot be sure to adhere to 'G1 truthful design' and therefore, they should refrain from using similarity since they cannot ensure that theirs is a responsible use (as noted, after explaining all three guidelines, I will discuss how to apply them together).

The second design guideline is concerned with ensuring a sufficiently symmetrical influence relationship between agent and user, leaving users substantial control over their mental states and behaviour (see Chapter 4). There are a few related reasons to worry that the use of similarity in persuasive agents could result in agents with an influence potential that endangers a user's substantial control. First, on the dimension of individual types of similarity, agents do not face certain natural human limits. For example, humans cannot morph their face to the face of their conversation partner, whereas avatars can be designed to do so, even if that might initially be a formidable design challenge. Also, characteristics such as the gender of the artificial persuasive agent are extremely easily adapted to match the user. Second, in particular, the option to *combine* several types of similarity in the design of a persuasive agent raises the concern that they might become *too* influential. Imagine a digital agent that combines all the ways in which it could be similar to you that have been discussed so far: it mimics you in several ways, it presents itself as a member of the same group and the same gender, it morphs its face to yours, it expresses similar preferences to yours, etc. Potentially, this agent could be extremely influential, although so far this remains an empirical question. Also, the attempt might backfire and cause reactance from users. Nevertheless, it is useful to ethically reflect on the option in advance.

With respect to this 'multi-similar' agent, the ethical concern cannot be discharged by noting that there are analogous human-human cases. The fact that there are humans, e.g., some used-car sellers, who are also highly influential on people whom they encounter through a clever and extensive use of similarity, is just as problematic: it leads to an overly asymmetrical influence relation, endangering the client's substantial control over her purchasing behaviour (cf. Cialdini: 2006). This type of car seller effectively replaces the consumer's judgement with his own, which amounts to manipulation (Sher: 2011). The existence of such salesmen is reflected in laws that provide cooling-down periods for customers of door-to-door sellers and for large purchases in general, allowing consumers to nullify their purchases.

Thus, the aim of the second guideline is to prevent the excessive use of similarity in persuasive technology:

Design guideline 2: the use of similarity in PT should allow for a sufficiently symmetrical influence relation between PT and the user ('G2 influence symmetry').

Of course, there is a great deal of room and a substantial need for judgment with regard to the application of this guideline. Individual users will vary widely (see also 5.5 below). However, the very same designers and deployers who have the knowledge to design similarity influence in persuasive agents are thereby also in the best position to apply 'G2 Influence symmetry' (and the other guidelines).

So far, this discussion has been concerned with the influence of persuasive technologies on users, but there are also ways in which a user can influence a PT, enhancing symmetry. One very important issue is the freedom to opt out, to decide to stop using the PT and to turn off the PT. One example would be to have the option to navigate through a web-shop without having to interact with a digital sales agent. Of course, if reaching important goals is only possible by using a certain PT, this opt-out is of limited value. In such a case, the user is better served with the option to change the settings of the PT.

A third guideline governs the use of similarity to enhance the usability of PT. Instead of designing for maximal influence on users, a designer might simply be concerned with developing a usable product. Perhaps some types of similarity must be incorporated to avoid ending up with an agent with whom users refuse to interact. Imagine a digital agent that always answers questions from users immediately or after four seconds. It could be that such response times are experienced as inappropriate, unnatural, or in some other way that has a negative effect on usability; the agent might just appear socially awkward. In the case of designing for usability, intentions to manoeuvre are absent, and the charge of manipulation does not hold. Therefore, we must distinguish between using similarity to design for usability and using similarity to design for maximal persuasiveness.

Nielsen (1993, 23) has defined usability in the context of human-computer interaction as involving five components: learnability (how easily users can learn to use the system), efficiency (how quickly users can perform tasks), memorability (how easily users "re-establish proficiency" after a period of non-use), errors (the amount and severity of user errors and ease of recovery), and satisfaction (whether it is pleasant to use the design).

For several of these components of usability, similarity might help to better realize them in the design. Programming agent response times to be fixed or random might feel so unnatural to users that satisfaction and efficiency are sharply decreased; confusion may even lead to more errors. So, it might be that some degree of mimicry is necessary to achieve a sufficient level of user satisfaction. In addition, there is a clear distinction between programming an agent to display some small degree of mimicry, just enough to ensure usability, and programming a larger degree to ensure maximal persuasiveness.

Therefore, my argument is that the employment of similarity to design for usability is non-manipulative since designers' intentions to exert a controlling form of influence on users are absent. However, in distinguishing between designing for usability and for persuasiveness, reference to designer intentions should be avoided. First, in practice it will often be difficult to assess which intentions guided a designer. Second and more importantly, what matters most for users is not what designers intend, but how an interacting PT 'treats them'. They deal with the product itself and not with its designer. Therefore, users may be strongly displeased if a PT exerts a strong influence based on similarity, even if its designer sincerely incorporated similarity to enhance usability.

The following design guideline is a criterion that avoids problematic reference to designer intentions.

Design guideline 3: If similarity is employed to design for usability, then that is a good reason for incorporating ('G3 design for usability').

The advantage of this formulation is that given the relevant expertise, it does not matter who performs the assessment of whether the guideline is met, and we need to know nothing about the designer's actual intentions. Note that this guideline governs only design for usability and is insufficient for determining whether a given use of similarity is morally justified.

After discussing G1-3 separately, I will now explain how the three formulated design guidelines function together in ways that rule out a manipulative use of similarity in PT. 'G1: truthful design' and 'G3 design for usability' describe justifying conditions or justifying grounds for making PTs similar to users, whereas 'G2 influence symmetry' sets a definite limit on these justified uses. Thus, even if designers combine two types of similarities in a way that each use meets 'G1: truthful design', if their combined effect leads to a PT that is too influential on its users, such design is ruled out by 'G2 influence symmetry'. With regard to 'G3 design for usability', it is at least conceptually possible that the incorporation of similarity in PT in a way that meets G3 simultaneously leads to a greatly enhanced persuasive potential. This concern is especially urgent in the case of designing PTs. In everyday human-human interaction, it is generally the easiest to interact with socially intelligent persons (the human equivalent of 'usability'), who consequently are, on average, more influential. The same may well apply to PTs: those which are by virtue of their artificial social intelligence most usable might also be most influential. Therefore, employment of similarity for usability must be limited by guideline 'G2 influence symmetry'.

Finally, every use of similarity for usability will also lead to user expectations with regard to how the persuasive agent is designed to interact with them. Therefore, for every use of similarity that is supported by 'G3 design for usability' designers should simultaneously attempt to conform to 'G1: truthful design'. In cases where that is not possible or where it is unreasonably complicated, they might still employ similarity for usability, provided 'G2 influence symmetry' is met.

5.5. Objections

Along the way, we came across a few lines of argument that might be objected to, some of which I will discuss. First, it might be objected that making PT similar to users to *induce* trust with technological means is always wrong because it does not provide users with adequate grounds for trusting agents. Thus, instead of using similarity to enhance user trust in persuasive agents, these agents should provide users with "sound evidence of their trustworthiness".[78] For example, they should make clear in general terms what users can and cannot expect from them, how their recommendations are related to user input of preferences, which company they represent, and so on.

This objection, however, fails to acknowledge that different types of evidence of trustworthiness may be at stake. Making persuasive social agents similar to users does not provide users with explicit information about agents' trustworthiness. However, if the PT is designed in accordance with design guideline 1 (i.e., the similarity truthfully represents how agents are designed to interact with users), similarity is a *reliable indicator of trustworthiness*. As noted in section 5.2, this works in much the same way in which similarity often induces trust between humans. Likewise, in human-human interaction, similarity is not itself the ground for trustworthiness but generally is a reliable indicator of such grounds, such as the fact that my conversation partner likes me or has a cooperative attitude towards me. Although it is relevant to ask whether designers should

[78] This objection assumes "evidentialism about trust and trustworthiness: the idea that trust should be based on sound evidence of trustworthiness". See for discussion (Nickel: 2011). See (Nickel et al.: 2010) for a clarification of how the notions of trust in technology and trustworthy technology can make sense.

also design their PTs to provide more explicit evidence of their trustworthiness, this is an independent question.

A second objection to this chapter's approach might be that it takes human-human interaction as a natural and therefore morally unproblematic baseline. Thus, it seems committed to some kind of is-ought fallacy and fails to be sufficiently sensitive towards what might go wrong in human-human interaction. The fact that some very common human behaviour is unreflective and automatic does not rule out the possibility that it is manipulative and moreover, that we can be responsible for such behaviour. I agree with all these observations.

My approach, however, does not take everyday social interaction between humans (whether it has a more deliberate and reflective or a more automatic and unconscious nature) as an always morally unproblematic starting point. Rather, we should carefully conduct our investigations on a case-by-case basis. Using the mimicry example, from social psychological investigation we know that in the paradigmatic cases, mimicry serves social interaction from which all partners benefit. Thus, the common good is served in ways that are non-manipulative. However, intentional mimicking between humans, while not the paradigm, remains relatively common and is more problematic. By carefully analysing when mimicking is and is not morally problematic and why, we better understand how to use mimicking responsibly in persuasive agents. More generally, the functions of some types of similarity-based influence might confer only a mixed blessing. For example, although group favouritism may have been indispensable to humanity in its earlier stages, seems also connected to all kinds of implicit biases from which vulnerable groups in society suffer. This should be reason to be very careful with the incorporation of similarity based on group membership.

Here, we encounter one advantage of artificial social agents over humans. Unlike us, they are not bound by inevitable psychological make-up, and thus designers have more control over responsible use. Of course, in this area, making accurate judgments will be difficult, value-laden and dependent on one's broader worldview, and therefore essentially contested. However, perhaps here we should say that in the event of doubt, designers should not incorporate similarity influences in persuasive agents. In any case, the approach taken in this chapter includes criticism of natural social influence and does not commit an is-ought fallacy.

A third but related objection states that all this places social psychologists in the relatively paternalistic position of determining the purposes and functions of various forms of similarity. Consequently, they also determine which benefits persuasive artificial agents should be designed to deliver to users.[79] Social psychologists indeed play a crucial role, but it is less clear that paternalism is involved. For paternalism to be involved, there must be interference with the will of the users for their own good, to which they do not consent (Dworkin: 2016). However, even if users place more trust in a digital agent, they remain free to determine the specific manner in which they cooperate with the agent, that is, whether or not to buy something, whether or not to ask for further medical advice, and so on. As with human interactions, people have the ability to reach their decisions based on all the reasons they find relevant in the situation. Thus, it is unclear how designers (and behind them, social psychologists) interfere with the will of users because providing users with the benefit of a trustworthy persuasive agent does not interfere with their will.

A fourth potential objection targets the assumption that users are sufficiently uniform in their psychological make-up and their responses to persuasive artificial agents for designers to be able to make responsible use of similarity influences. Given the wide diversity among humans in terms of their personality, intelligence, experience, socio-cultural background, etc., the accuracy of this assumption might be questioned. My response must be brief and tentative, but it centres around the idea that with great power comes great responsibility. Artificial social agents are sufficiently intelligent to adapt so that they become similar to individual users. Therefore, we might hypothesize that the agent will also be intelligent enough to infer how the user's psychological makeup and her responses differ from the average *and adapt to these characteristics*. Thus, the persuasive agent likely could be designed to be able to adapt to individual users, at least to some extent. If so, we might say that it can be designed to operate on the basis of a 'personalized ethics'.

Finally, some may still object that my account is too permissive with respect to the use of similarity influence. If so, it would be good to realize my account's many restrictions on the use of similarity. The use of similarity results in the acquisition of new obligations that may be so strict that the designers may prefer to abandon the use of similarity altogether, or use it sparsely—for example, merely for purposes of usability. The three guidelines rule out the strategic uses of similarity from which designers and deployers, but not users, benefit.

[79] I thank Eric Schliesser for pressing this objection.

5.6. Concluding remarks

In this chapter, I have first explained the psychology of similarity-based influence. This revealed that the concern about manipulation arises from the fact that human users tend to respond unreflectively and automatically ('mindlessly') to social behaviour by persuasive agents in general and to similarity influence in particular. From the perspective of Habermas's ethics of communication, I argued that we primarily have reason to be concerned about manipulation if the validity claim to truthfulness (raised by the persuasive agents' communication) is not satisfied—that is, if the similarity incorporated into the design and communicated to the human user does not truthfully represent or indicate how it is designed to interact with users. Thus, if similarity induces trust in users, the persuasive agent should be trustworthy.

On a concluding note, our society should ask whether we are willing to use the services of all kinds of artificial social agents that sometimes will aim to persuade us, just like fellow humans. If we prefer their help in web-shops, as social robots caring for the elderly, as digital medical assistants, or as online teachers, we probably need to accept *some* types of similarity influence for these agents to be sufficiently social (and sufficiently persuasive to start using them in the first place). Interacting with a completely non-mimicking agent might be so awkward that it would be of no help at all. To ensure the responsible use of similarity influence by the sort of artificial social agents just mentioned, I have developed the three design guidelines discussed above ('G1 truthful design', 'G2 influence symmetry', and 'G3 design for usability').

Finally, the approach taken in this chapter will probably, *ceteris paribus*, also be valuable to apply to other sources of social influence, as many other forms of social influence can be construed as part of the communicative interaction between humans and persuasive agents. For example, think of giving social praise (Kaptein et al.: 2011). Extending the argument of this chapter in such directions would be an interesting avenue for further study.

6 Scanlon on acceptable risk-imposition

6.1. Introduction

An important class of persuasive technology concerns persuasive technology that is employed for risk reduction. Think of the Tesla example in the introduction. In the Tesla, persuasive technology is employed to make sure that drivers keep their hands on the steering wheel and remain focused on their driving task. Another example concerns persuasive gaming to train airplane passengers for emergencies (Chittaro: 2012). The goal of these types of persuasive technologies is to reduce a users' risk to a more acceptable level. Other persuasive technologies are also or primarily designed to reduce certain risks imposed on people other than the user of the PT in question. One example is Intelligent Speed Adaptation, that warns drivers any time they exceed the maximum speed. If we conclude that driving a car generates unacceptable risk levels for pedestrians, cyclists and drivers, our government may make it mandatory to use Intelligent Speed Adaptation (see Chapter 7 below).[80]

Looking from a more general perspective, the question of the conditions of acceptable risk imposition is important because of the large and crucial role of technology in society. As a result of technology, it has become increasingly a matter of human choice and policy which risks are generated and how the benefits and risk burdens are distributed over different citizens and groups of citizens.

To determine whether risk reduction is desirable or even necessary, through persuasive technology or some other means, we need to create a plausible moral framework for acceptable risk. However, the question of acceptable risk is complicated and fundamental. On the one hand, even though virtually all actions bring with them some risk of harm, we often need to perform them to obtain

[80] There is another relation between risk ethics and persuasive technology, which will not be treated in this chapter. The use of persuasive technology may in some cases generate risk. For example, ill-designed eco-feedback systems in cars may distract drivers, causing accidents. Or, drivers may brake too late in order to save fuel. As another example, in case of badly functioning persuasive technology designed to help smokers quit, we should not just say "it can't hurt to try". On the contrary, disappointed users may refuse a next attempt for years.

important benefits and to live a worthwhile life. On the other hand, our fellow citizens have a strong entitlement to remain free from harm. This raises the question of how the benefits and burdens of risk-generating actions should be weighed in order to assess the moral acceptability of these actions. Consequentialist approaches allow for interpersonal aggregation, in which small benefits to many agents can be added up to justify significant risks to fewer agents. Thus, these approaches, or at least the less sophisticated ones, allow for the imposition of significant risk burdens on some individuals while the benefits of the risk-generating actions go to others, given that the net utility is maximal. Deontologists object to this approach, arguing that interpersonal aggregation either violates fundamental individual rights or fails to respect the 'separateness of persons' (Rawls: 1971). However, deontological approaches create the danger of imposing such stringent conditions on the acceptable imposition of risk that virtually no action is acceptable. The rights of individuals to remain free from harm, especially from harm to one's bodily integrity, are so strong that they are taken to entail a right to also remain free from the risk of harm. Strict limits for risk-generating agency follow that would bring normal life to a standstill: the so-called 'problem of paralysis' (Altham: 1983; Hansson: 2003; Hayenhjelm & Wolff: 2012). The attractiveness of Scanlon's contractualism lies in its attempt to avoid both interpersonal aggregation and the problem of paralysis (Scanlon: 2000).

In this chapter, I will argue that Scanlon's contractualism is an attractive moral framework for determining acceptable risk imposition because he can indeed avoid appeal to interpersonal aggregation. However, I will also argue that it is not clear whether he ultimately can avoid the problem of paralysis. Some critics hold that Scanlon's contractualism clearly cannot avoid this problem (Ashford: 2003; Fried: 2012; Norcross: 2002). The reason for this suggested inability can be traced directly to the core of Scanlon's contractualism: 'justifiability to each person concerned'. According to Scanlon, the principle for regulating risk imposition should be one that no one can reasonably reject (Scanlon: 2000). Deciding whether such a principle cannot reasonably be rejected involves pairwise comparison of an individual's reasons to object to the principle. That principle is justifiable to which the smallest individual complaint can be raised.

Crucially, Scanlon holds that an individual can complain on the basis of the full magnitude of harm *if the risk were to materialize*. Thus, his theory rejects discounting the magnitude of the harm by the probability of its occurrence. For example, a pedestrian could object to a risk-regulating principle that allows cars to drive 50 km/h in a residential area. This is because driving 50 km/h could lead to her death, regardless of the probability that it would happen. Scanlon argues that the probability of a lethal accident nevertheless is relevant, but for purposes of determining how much precaution the driver should take when driving 50 km/h in the residential area (op cit., 209). However, even if all drivers have taken due precautions, some pedestrians will be killed. A potential victim's complaint about the principle *permitting* driving will always be stronger than a driver's about the principle *prohibiting* driving. Thus, contractualism seems to support principles that prohibit any action that imposes risk of death, leading to extreme demandingness (Ashford: 2003) or 'moral gridlock' (Fried: 2012). Ashford and Fried both argue that to avoid this highly implausible implication, contractualism *must* appeal to some form of inter-personal aggregation.

I will argue, however, that these critics have not sufficiently appreciated Scanlon's employment of the notion of *intra*-personal aggregation of risks and benefits, which enables him to avoid this implication to a large extent. Intra-personal aggregation here refers to summing up all the (agency) benefits and (risk) burdens of a given risk-regulating principle over the course of an individual's lifetime (Scanlon: 2000, 237). If my argument is successful, this is an important result because Scanlon's contractualism is defined in opposition to inter-personal aggregation.[81] Nevertheless, I will show that his contractualism leads to relatively stringent constraints on acceptable risk imposition, which many will still find implausible and view as leading to a state close to paralysis.

In section 6.2, I will first sketch the main contours of Scanlon's contractualism. In section 6.3, I will discuss Scanlon's views on acceptable risk, including his appeal to intra-personal aggregation. I then discuss the role of intra-personal aggregation in more depth. I argue that intra-personal aggregation enables us to principally avoid moral gridlock, but nonetheless entails fairly stringent constraints on acceptable risks (6.4). These constraints may first appear unreasonably strict, but there are several mitigating considerations (6.5). I conclude by drawing three significant implications for how societies ought to govern risk (6.6). Although these considerations give guidance for reflection on persuasive technology and risk imposition, the mitigating considerations of section 6.5 render Scanlonian contractualism less than fully determinate.

[81] For, the comparison of individual objections to principles for the regulation of behavior is pairwise: there is "no individual who enjoys" "the sum of the smaller benefits" that could justify the greater burden of another individual (Scanlon: 2000, p229-230).

Therefore, there is a brief mention of additional guiding considerations that are still necessary. Together with the three implications, these considerations are subsequently taken up in Chapter 7 on Intelligent Speed Adaptation as a means for reducing driving risks.

6.2. Scanlon's contractualism

Scanlon's contractualism seeks to find moral principles that no one could reasonably reject. In finding such principles, Scanlon is concerned with the domain of morality that regards "what we owe to each other", thus, with acting rightly or wrongly towards others. The 'basic formula' of his contractualism is: "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced general agreement" (153). Crucially for Scanlon, this presupposes *morally motivated* persons who wish to justify themselves to each individual person concerned. This core feature of Scanlon's contractualism is Kantian in the sense that in this way, persons are treated as ends in themselves.[82] Scanlon's contractualism should be sharply distinguished from forms of contractarianism that concern rationally self-interested persons seeking (actual) agreement for mutual advantage (cf. Ashford & Mulgan: 2012).

Assessing both potential principles for the general regulation of behaviour and alternatives to these principles consists of pairwise comparing individuals' set of reasons for rejection. Thus, when considering whether it is wrong to perform a certain act in certain circumstances, we need to consider both principles that permit the act and principles that prohibit the act. We need to assess the burdens that principles allowing the act would place on different individuals affected by the act. Also, we need to assess the burdens that principles prohibiting the act would place on the agents. We then must compare the reasons of each individual (the agent and each person affected by the act) to reject the several possible principles. In Asfhord's words, "we converge on a principle that no one can reasonably reject when the individually strongest objection to it is as small as possible" (Ashford: 2003, 280).

[82] Scanlon, however, is eager to emphasize the distinctiveness of his contractualism with respect to Kantianism. See (T. M. Scanlon: 2111).

Scanlon's well-known discussion of a 'rescue principle' will serve to illustrate this pairwise comparison of reasons for rejecting principles (op cit., 235). He argues a principle that requires that "if one can save a person from serious pain and injury at the cost of inconveniencing others [...], then one must do so no matter how numerous these others may be". His example is Jones working in the transmitter room of a television station, on whom electrical equipment has fallen, causing extremely painful electrical shocks (which, however, cause no permanent damage). The only way to rescue Jones is to briefly interrupt the broadcast of a World Cup match, which millions of people are watching. The above rescue principle would require interrupting the broadcast to save Jones. The reason for each of the watching people to reject this rescue principle is the interruption of their pleasurable experience of watching the World Cup. The reason for Jones to reject alternative principles that permit the broadcast to go on is clearly stronger: having to experience extreme pain for the remainder of the broadcast. It is clear that Jones' reason for rejecting these alternative principles is much stronger than each individual viewer's reason for rejecting the rescue principle.

A plurality of considerations can figure as a reason for the reasonable rejection of a principle. Of foremost importance for Scanlon are consequences for individual well-being. These consequences are defined in terms of a principle's impact on each individual's relatively general interests in connection to his well-being, such as health and life (Cf. Ashford, 2003, 277). In this way, Scanlon's theory captures the fundamental intuition that gives so much plausibility to utilitarianism: the manner in which our acts affect the well-being of others is morally important (Scanlon: 1982). Unlike utilitarianism, however, Scanlon's contractualism only acknowledges individual costs to well-being as reasons for objection, not the aggregated costs to a number of different persons. In addition to well-being, entitlements can also serve as reasons for rejecting principles. One could, for example, think of entitlements stemming from obligations created when others make promises. Furthermore, if principles treat persons differently for arbitrary reasons, this would be unfair, which is also a ground for reasonable rejection.

Importantly, the presence of options for voluntary choice for the individuals that are affected by a principle, will shape the way in which these above considerations enter the contractualist reasoning. If individuals can choose to act in ways that reduce the burden that the principle imposes on them, then the force of their objection to this principle made on the basis of that burden is diminished. Scanlon denotes this idea with the label 'responsibility' (op cit., 249), and it is clearly relevant to the determination of risk-regulating principles. If the cyclist in our earlier example can choose between a cycle track and the road, she cannot object to principles permitting driving based on the risk burden that results from road cycling together with cars. In contrast, pedestrians have no other choice than to use the pavement and occasionally cross streets.

Scanlon does not appeal to a Rawlsian veil of ignorance. Instead, his version of contractualism turns on reasons for rejection that arise from the point of view of individuals with full knowledge of their situation. In other words, he grants individuals knowledge from their own standpoints, though characterized in suitably generic terms. The reasons for objection that arise from those standpoints should be generic, reasons that individuals in that situation typically would have. The rationale for abstracting to generic reasons is epistemological. Since we cannot know which specific individuals will be affected by principles and how, we can only appeal to "commonly available information about what people have reason to want" (op cit., 204). However, even if some proposed principle would affect only one person, that person could still proffer generic reasons for refusing to accept that principle. This is the case because these reasons are generic: any person in his situation would have these reasons to object.

When considering principles, we do not only consider the costs of permitting and prohibiting single actions to agents and those affected. Instead, we consider how the general permission or prohibition of the widespread performance of such acts would affect individuals over a longer period of time. Because we shape our life partly in response to what is prohibited and what is permitted, this impact on individuals goes beyond the direct consequences of those acts. For example, the principle that prohibits inflicting bodily harm on me without my consent enables me to 'navigate more confidently through social space'. However, these more general costs of permitting or prohibiting classes of actions are costs to individuals; they only enter contractualist reasoning as an individual's generic reasons to object to principles. This brief summary of Scanlon's contractualism should enable a more detailed explanation of how he views the permissibility of imposing risk.

6.3. Scanlon on risk-regulating principles

In this section, I will reconstruct how Scanlon addresses the issue of the principles that regulate risk-imposition. I think it is fair to say that he does not provide a systematic and comprehensive treatment of this issue, as he pays only brief attention to it, spread over different places. However, what he has to say is interesting enough to merit further investigation. Furthermore, since the issue of acceptable risk-imposition is one of central importance, his contractualism (and any ethical theory) must harbour the resources to treat it satisfactorily.

Comparing burdens in cases of harm that is certain is straightforward, at least sometimes. Cases of risk, however, in which harm is uncertain, are complicated for contractualism. This is because they pose the question of whether the objection of those exposed to a certain risk of harm should be based on the full magnitude of the harm in question. On the one hand, if the probability of harm is not considered, then the reasons to reject principles that allow behaviour that brings some risk of great harm are as strong as the reasons to reject principles that allow behaviour that is certain to cause the same harm. On the other hand, if the harm is discounted by its probability, unfortunate implications result as well. Scanlon gives the example of a lottery that would assign a few people the role of subjects of painful and dangerous medical experiments that aim to benefit many others. The fact that each citizen has only a very small chance of becoming a victim of this lottery does not, according to Scanlon, diminish her complaint; in rejecting principles that license such experiments, she can appeal to the full burden of being an experimental subject.

What is at stake here is the choice between an *ex ante* and an *ex post* version of contractualism.[83] These versions differ in their specification of the epistemological position in which people seek principles that no one could reasonably reject. '*Ex ante'* means 'before the event, and *'ex post'* means 'after the event', or 'afterwards'. *Ex ante* versions of contractualism determine which principles no one can reasonably reject based on what individuals can know beforehand. *Ex post* versions of contractualism grant individuals hypothetical knowledge of how such principles would *in fact* shape their life.

The manner in which this epistemological position is specified does a great deal of moral work in determining principles of acceptable risk. According to *ex ante* contractualism, individuals have no knowledge of how the various alterna-

[83] For extensive discussion, see (B. H. Fried: 2012; James: 2012; Lenman: 2008)

tive risk-regulating principles considered will affect their lives. They do not know whether they will be seriously harmed as a result of some risk-generating activity permitted by a principle, but at most have a more or less reliable estimate of the probability of becoming harmed. Any possible individual complaint against principles must appeal to expected outcomes, which involves discounting harms and benefits by their likelihoods.

In contrast, *ex post* contractualism allows individuals to object on the grounds of how proposed principles will actually affect them. For example, any citizen can object to principles permitting Scanlon's medical experiment based on the pain and danger caused by being a subject. Potential victims of such experiments are allowed, in Fried's terminology, to "peek ahead" and see how they will be affected by a principle permitting these experiments (Fried: 2012, 43). In other words, even though we must agree on principles at the point in time before the activities allowed by these principles cause real harms, *ex post* contractualism grants individuals the option to object based on the occurrence of bad outcomes after principles are adopted. Therefore, because in cases of risk an individual by definition cannot know whether these bad outcomes will happen to him, it must be that he is allowed to appeal to *possible* bad outcomes *as if* they would happen to him with certainty.

Scanlon seems to avoid a clear choice between discounting or not and on a closely related note, between *ex ante* and *ex post* contractualism. He comes up with an alternative method of accounting for the relevance of the probability of harm:

> the probability that a form of conduct will cause harm can be relevant not as a factor diminishing the 'complaint' of the affected parties (discounting the harm by the likelihood of their suffering it) but rather as an indicator of the care that the agent has to take to avoid causing harm (op cit., 209).

The idea here is that greater possible harm and greater probability of its actual occurrence are both reasons for the agent to take more stringent precautions. To use the above driving example again, the greater the probability of lethally hitting a pedestrian, the greater the driver's duties to take precaution. For example, she could drive slower, in a car that is designed for pedestrian safety.

After arguing for this role of probability, Scanlon goes on to explain that it "would be too confining" (op cit.: 209) to demand more of agents than taking reasonable care and precaution. Clearly, provided we have taken such reasonable precaution, we cannot forego all risk-generating actions since this would be too burdensome for agents. Therefore, a principle prohibiting all risky actions can reasonably be rejected by agents.

This notion of reasonable precaution is plausible and worthwhile, but as long as Scanlon does not allow for discounting, this method of factoring in the probability of harm will not be sufficient to provide room for risk-imposing actions. Consider Scanlon's own example of air travel (op cit., 209). We think air travel is permissible, even though planes may crash and cause the death of people on the ground. We think so, Scanlon holds, because reasonable care is taken, and forgoing air travel would be too confining. However, the set-up of Scanlon's contractualism does not allow him this move since as we have seen, it explicitly holds that the fact that a harm is very unlikely to materialize does not diminish the complaint of the potential victim. Recall that Scanlon needs to claim this in order to be able to rule out as morally unacceptable cases like the medical lottery (see above). However, as Ashford has forcefully argued, if we compare the burden of being killed by a plane crashing on the ground with the burden of having to forego air travel, it is clear that potential victims have the strongest reason to object to principles allowing air-travel (2003, 298-99). Thus, it seems that Scanlon indeed does face the objection of moral gridlock: against any risk-generating action, a potential victim can raise a moral veto.

However, in the second main locus, in which Scanlon addresses principles regulating actions that impose a risk of harm (op cit., 235-238), he provides the resources to respond to this criticism. In the following passage (which is worth quoting in full), he recognizes the force of the objections by victims such as those of plane crashes:

> Suppose, then, that we are considering a principle that allows projects to proceed, even though they involve risk of serious harm to some, provided that a certain level of care has been taken to reduce these risks. It is obvious what the generic reason would be for rejecting such a principle from the standpoint of someone who is seriously injured despite the precautions that have been taken. On the other side, however, those who would benefit, directly or indirectly, from the many activities that the principle would permit may have good generic reason to object to a more stringent requirement. In meeting the level of care demanded by the principle, they might argue, they have done enough to protect others from harm. Refusing to allow activities that meet this level of care would, they could claim, impose unacceptable constraint on their lives (op cit., 236-7).

Scanlon clearly recognizes that prioritizing the burden of the victims of risky activities that result in harm would "impose unacceptable constraints" on the lives of agents. His way to avoid this implication is by stipulating that those who are the potential victims of a variety of risk-imposing actions *are the very same individuals* who stand to benefit from being allowed to perform these actions:

> The contractualist argument I have just stated includes a form of aggregation, but it is aggregation *within* each person's life, summing up all the ways in which a principle demanding a certain level of care would constrain that life, rather than aggregation *across* lives, adding up the costs or benefits to different individuals (op cit., p 237, emphasis in original).

As we see, Scanlon explicitly qualifies the type of aggregation that he needs to avoid moral gridlock as intra-personal (as the opposite of inter-personal) aggregation.

This scheme of intra-personal justification allows individuals to perform a whole range of actions, which may differ substantially between individuals. Scanlon's contractualism searches for principles for the "general regulation" (op cit., 153) of behaviour and does not require a separate principle for regulating each risky activity. Thus, it is not required that some person perform a certain action *herself* and benefit from performing it for another person to be permitted to perform that action. The idea is that people 'pool' their risks.[84] Each both is burdened by the many different risks imposed on her by others and benefits by being allowed to engage in many risk-imposing activities herself (provided reasonable precaution is taken).

In this way, Scanlon succeeds in warding off the charge of moral gridlock. Over a lifetime, individuals each benefit intra-personally from principles for the regulation of risk imposition that permit mutual risk imposition. Therefore, it appears that no one can reasonably reject such principles.

However, this appeal to intrapersonal aggregation commits Scanlon to an *ex ante* contractualism. In an *ex post* version, an individual could still appeal to the possibility of being killed as a cost, and intrapersonal aggregation involving this cost would still block principles permitting risks that might materialize into lethal harm. Thus, the pedestrian could still object to principles allowing cars to drive 50 km/h on the ground that such a practice might kill him. Because of the inclusion of death, the costs of principles permitting mutual risk imposition would always outweigh any agency benefit.

[84] For an interesting treatment of the concept of a risk pool, see (C. Fried: 1970). See also (S. O. Hansson: 2003).

Whereas *ex ante* contractualism seems to be the only version that is capable of avoiding the problem of moral gridlock, it clearly introduces new problems. *Ex ante* contractualism does not enable Scanlon to block medical lotteries based on the full magnitude of harm to the subjects of the medical experiments. Along the lines of some version of *ex ante* average utilitarianism (Harsanyi: 1953, 1982), it could be *ex ante* beneficial to each of us to accept Scanlon's medical lotteries. This would be so as long as the expected utility of the medical experiments, in terms of lives saved and better health for many, exceeds the expected disutility of the unhappy few who are chosen as subjects in the medical experiments.

It will be clear that developing plausible contractualist constraints on what would be *ex ante* justifiable to each is a serious challenge. In any case, attempting it is far beyond the scope of this chapter and I will leave the issue of whether it can be done as an open question (for an illuminating attempt, see James: 2012).

In the next section, I will take a closer look at the role of intra-personal aggregation in Scanlonian contractualism and show how it leads to stringent restrictions on acceptable risk imposition.

6.4. Interpreting Scanlon's appeal to intra-personal aggregation

Upon closer inspection, it turns out that individuals do not benefit from a given set of risk-regulating principles to the same degree. This is problematic for Scanlon's contractualism because, as we will see, this implies very stringent risk-regulating principles. The outcomes of the summing up and weighing of the burdens and benefits that over the course of one's lifetime would result under some set of risk-regulating principles, will differ significantly between individuals. Some will fare much better than others. In addition, individuals' judgments about when risk-regulating principles become too constraining, aggregated over the course of one's life, obviously are widely different. Consequently, it is not at all immediately obvious how their individual objections determine the set of risk-regulating principles that no one could reasonably reject.

To explain different ways in which these individual objections could determine the set of risk-regulating principles that no one could reasonably reject, I will compare the following two individuals living in the same neighbourhood. The first is a very wealthy CEO. He drives 80,000 km per year in a 2003 diesel-fuelled SUV, and his hobby is sports flying. In general, his level of consumption is significantly above average. The second is a 60-year-old nurse, who has a modal income and travels exclusively by public transport. She also suffers from a heart disease that makes her more vulnerable to polluted air. The CEO and the nurse impose a significantly different level of risk on each other. The CEO imposes on the nurse substantial safety risks caused by the mass and construction of his SUV (Husak: 2004; Vanderheiden: 2006) and severe health risks caused by the emission of black carbon (Loomis et al.: 2013),[85] for which especially old diesel cars are notorious. However, the nurse does not impose these risks on the CEO. The CEO might die from a collision with the train or bus used by the nurse, caused by malfunctioning warning systems, but the likelihood that the nurse will die from a collision with the CEO's SUV clearly is much greater. A moment of reflection on this example will reveal that in general, there is great variance in the levels of risk that individuals impose on each other.

Whereas the CEO would need a relatively liberal set of risk-regulating principles to be permitted to perform all of his risk-generating activities, the nurse could operate with a relatively strict set. The question is which of the two can raise the strongest objection to the strict and the liberal set of principles, respectively. Answering this question involves some interpretation of Scanlon's appeal to intra-personal aggregation. Any set of risk-regulating principles would allow certain activities, specify the level of precaution to be taken, and prohibit other actions. Intra-personal aggregation is the summing up, across an individual's life, the costs and benefits of living under that set of principles. According to Scanlon, the costs involve required precaution, prohibited actions, and risk exposure. On the side of the benefits are the many actions the set would permit, including indirect benefits from these actions.

Comparison of the outcomes of different individuals' intra-personal aggregation should determine which set of risk-regulating principles no one could reasonably reject. Of course, many sets would benefit each individual in the sense that permission to act even at high costs in terms of risk-exposure and precaution is always better than moral gridlock. However, that is too weak. The question is which set of risk-regulating principles an individual will choose based on the intra-personal aggregation of costs and benefits alone. The nurse would prefer a relatively strict set that, roughly, allows her the agency benefits she needs to live her life at a relatively low cost of precaution and risk exposure. Under that strict set, the CEO would have the same agency benefits for roughly

[85] http://www.iarc.fr/en/media-centre/pr/2012/pdfs/pr213_E.pdf, accessed 06-11-2014

the same level of risk exposure (though he will value the two differently). However, the CEO would be willing to assume higher risks and take more precaution in return for more agency benefits.

Since their intra-personal weightings of costs and benefits lead the nurse and the CEO to opt for a strict and a liberal set, respectively, we need to know who can raise the strongest objection. The nurse's reason for rejecting the liberal set is obvious: it puts her under a significantly higher risk-burden than the strict set, for no additional benefit.[86] The CEO, however, would object to the strict set of principles by showing, in Scanlon's terms, the 'ways in which it would be constraining his life'. He could object based on the costs of having to forego many activities that make his life worthwhile. I will call these costs 'agency-costs'.

Determining who will face the greatest burden, then, raises familiar questions of interpersonal comparison of burdens and benefits. What is a greater burden: a higher level of risk exposure or a lack of permission for several activities that make one's life worthwhile? This seems difficult to assess. Could the nurse and the CEO meet somewhere in between, at an intermediately strict or liberal set of risk-regulating principles, each accepting some deviation from their ideal trade between costs and benefits?

Before trying to answer this tough question, however, there is reason to ask whether agency costs should be permitted to enter the contractualist reasoning of 'reasonable rejection' in the first place?[87] A positive answer seems to presuppose what must be argued from the contractualist perspective, viz., that it is not morally wrong to perform actions that impose some risk on others. For comparison, consider consequentialism and deontology or rights-based theory. In a consequentialist framework, agency is justified insofar as it results in the best

[86] In the next section I will discuss ways in which she might benefit indirectly from others being allowed to generate more risk than she herself does. But I believe that these would not affect the basic argument here.

[87] I find it not so clear what Scanlon's position is on this question. On the one hand, instead of granting us an entitlement to risk-generating agency at the outset, Scanlon's reasoning seems to aim at explaining how, given the possibility of catastrophically bad outcomes it is nonetheless permissible to impose risks on each other. On the other hand, from his repeated talk in terms of how principles would be constraining a life it seems clear that he regards disallowed actions as a cost that is a basis for reasonable rejection. I agree that they are a cost, but it does *not* automatically follow that these costs also have to count in the intrapersonal cost-benefit analysis that determines which set of principles no one can reasonably reject.

consequences for all concerned. In a rights-based theory, following Kant, individual agents are granted basic agency rights or freedom to act, provided other persons are given due care (Steigleder: 2012). The contractualist framework should show how moral reasoning, in which each individual is motivated to justify herself towards each other individual concerned, results in a set of principles that allows for risk-generating agency. However, if such an 'entitlement' towards risk-generating agency is allowed to enter the apparatus of determining which principles cannot reasonably be rejected, it seems that we have circular reasoning: the conclusion (room for agency) enters the reasoning that leads to that conclusion. The charge of circularity is familiar with respect to contractualism, and here we encounter another instance of potential circularity (cf. Ashford & Mulgan: 2012; Hooker: 2002; James: 2004).

It is instructive to reflect on how contractualist reasoning would proceed if agency costs were included. In that case, there clearly must be limits to the agency costs individuals can take up in their cost-benefit calculation. Otherwise, an individual with an extreme risk-generating way of life (the CEO) would register a very strong objection based on his high agency costs under a strict set of risk-regulating principles. It would clearly be unreasonable to grant the most extreme individual the largest complaint based on the fact that he would have the largest agency costs. Therefore, it is clear that any reasonable appeal to agency costs under a set of strict principles at least would have to be made *within the limits* of a certain *baseline of a reasonable array of activities necessary for leading a good life*. Within a society, we need to agree what level of agency should minimally be granted for individual citizens to live according to a 'reasonable plan of life' (cf. Rawls: 1971). Given the increasing number of global challenges, such as climate change, we will also need such agreement on a global level.

At this point, the challenge for contractualism is to provide a more or less principled way, following contractualist reasoning, to determine that reasonable baseline. On a practical level, this baseline could not be too far removed from what a society currently regards as reasonable. Otherwise, the theory would not be accepted and thus be inapplicable in practice and of little use. On a more fundamental, theoretical level, however, it seems unclear how contractualism could give much guidance here. The reason is that it must be determined at the outset what would qualify as a reasonable baseline of acceptable risk-generating agency. Only after that has been done can agency costs enter contractualist reasoning for determining which principles no one can reasonably reject. Thus, the determination of the baseline cannot be the outcome of applying the contractualist framework. I conclude that it is currently unclear how Scanlonian contractualism can appeal to agency costs in a clear and principled way. Therefore, I will next investigate the implications for acceptable risk-imposition that follow if such appeal is excluded.

Supposing that Scanlon indeed should not include agency costs in intra-personal aggregation, individuals such as the nurse determine the standards of acceptable risk. She could reasonably reject any set of risk-regulating principles more liberal than that allowing her own agency-generated risk imposition on others because any more liberal set is no longer intra-personally beneficial for her. The implication is that at least under this interpretation, contractualism leads to the adoption of strict risk-regulations, regardless of whether they maximize social welfare. Individuals such as the nurse function as what may be called 'risk-dictators': they determine the standard of acceptable risk for all other individuals.[88] That standard is highly restrictive. It only allows roughly equal levels of mutual risk imposition, and these levels are as low as the level of people like the nurse. At first sight, this level appears to be so low that some will regard it as leading to a state close to paralysis or moral gridlock. However, in the next section, I will discuss a few considerations that mitigate this implication considerably.

6.5. Mitigating considerations

This contractualist equality requirement on mutual levels of risk-imposition may appear so stringent as to be outright implausible. It may seem that the CEO in my example is under a moral obligation to give up most of the activities that make his life worthwhile for him. It is only in that way that he will manage not to impose more risk on the nurse than she does on him. However, there are some considerations that will lead to a more nuanced picture.

The crucial insight is that risk-regulating principles confer to us not only the direct benefits of the actions that we are allowed to perform but also the indirect benefits of the actions of others (Scanlon: 2000, 237). The CEO and the nurse both live in a society that provides many goods and services that benefit them in many ways. They receive education, they can hold jobs that provide an income to buy products they need, they receive health-care, and so on. The production and

[88] Thanks to Wybo Houkes for the term.

delivery of these goods and services inevitably generate risks. Insofar as the nurse and the CEO purchase and rely on these goods and services, they are implied in the risks associated with these goods and services imply. We could say that they *indirectly* impose risks on each other and other members of society.

Both direct and indirect risk-generating actions and practices must be regulated according to principles no one can reasonably reject. The nurse has reason not to reject principles that allow the risks associated with a functional society. Aggregation of their costs and benefits to her, intra-personally, will be beneficial on balance. To know the extent to which the CEO must reduce his risk-generating activities, we need to know the extent to which those activities contribute to or are essential to societal goods and services. For example, in Western societies, at least in the short term, without auto-mobility, society would come to a virtual stand-still. Additionally, chemical companies produce basic material for many products that all citizens use. On a more general level, modern societies have developed high-quality health services that benefit each citizen in terms of risk-reduction. However, this modernization process, including economic development, also creates various new risks.

Consequently, determining which risk-regulating principles no one could reasonably reject depends on many considerations and complicated empirical information. Therefore, Scanlon's contractualism does not enable the derivation of principles for acceptable risks in a straightforward and undisputed way. Nevertheless, it is possible to draw some plausible implications of contractualism. The equality requirement entails a moral obligation for risk-reduction up to the point that further reduction would increase indirect risks or undermine risk-reducing goods and services, for example, by threatening their funding. For example, whereas the CEO probably has no obligation to give up car-driving, he should exclude all trips for trivial purposes and buy another car that pollutes less and has a much better design for pedestrian safety. In addition, his chemical company should be subjected to stricter emission norms, given the stipulation that this is economically possible (which is not the same thing as having a positive cost-benefit ratio[89]). In general, the equality requirement entails a strong presumption in favour of risk reduction.

[89] This would be the case if the company has funding for additional precaution, but the expected health benefits, expressed in an economic metric, are less than the economic cost of the additional precaution.

Opinions on whether these implications are plausible or too limiting to allow persons to live the good life that they wish will differ. In any case, despite the equality requirement, Scanlon's contractualism clearly succeeds in avoiding the problem of moral gridlock or paralysis. It allows many risky actions to be performed. In contrast, a moral theory that forbids any risk-imposing action is plainly incoherent, and whether the equality requirement is too confining is a matter of reasonable debate.

Implausible or not, the equality requirement is a direct consequence of Scanlon's commitment to avoid interpersonal aggregation. Contractualism shares this notion of the 'separateness of persons' (Rawls) or what Nozick calls the 'inviolability of individuals' (Nozick: 2013, 31–33) with other deontological ethical theories. It is worthwhile to ask whether it is easier to uphold this notion in the domain of actions with consequences that are certain than in the domain of risk (actions with uncertain consequences). It may be less confining to forego sacrificing individuals for the sake of benefitting many others than to forego any risk-generating action that would lead to unequal levels of mutual risk-imposition (and thus benefit some at the expense of others).

However, the cumulative effects of violating the separateness of persons in the domain of risk may be grave. Consider again the problem of air-pollution. Traffic is a major source of the most dangerous type of particulate matter pollution.[90] Energy power plants are another source. In Europe, 1-3% of cardiopulmonary and 2-5 % of lung cancer deaths are attributable to particulate matter, with children, the elderly, and the diseased being especially vulnerable. There is no evidence of a safe level. Allowing citizens unlimited energy consumption and driving for both trivial and essential purposes thus significantly contributes to a significant risk of premature death for fellow citizens.

I take it to be a major advantage of Scanlon's contractualism that it requires to compare individual complaints to risk-regulating principles. Thus, for example, it asks us to compare the health risk-burden on the elderly nurse suffering from heart disease to the burden on the CEO of driving less in a newer car. But again, whereas our evaluations of its implications will differ, Scanlon's contractualism makes clear that a commitment against interpersonal aggregation in the

[90] http://www.euro.who.int/__data/assets/pdf_file/0006/189051/Health-effects-of-particulate-matter-final-Eng.pdf Accessed 07-11-14

domain of risk has far-reaching implications. I will elaborate these implications a bit further in the next section.

6.6. Concluding observations

I have argued that Scanlon's contractualism includes the resources to avoid the problem of paralysis or moral gridlock. It indeed provides a middle ground between overly permissive consequentialist approaches on the one hand and overly restrictive versions of deontology or rights-based approaches on the other hand. Scanlon's essential move is to appeal to intra-personal aggregation of risks and benefits to justify risk-generating actions. Risk-regulating principles can be justified that allow one person to perform a whole range of actions that bring some risks to others since others are allowed to do the same. No one could reasonably reject such principles because overall, they are beneficial to each person throughout her life.

I have briefly argued that agency costs should not be included in the contractualist reasoning on which set of risk-regulating principles cannot reasonably rejected. If agency costs are excluded, Scanlon's appeal to intra-personal aggregation only works if the levels of risk that people impose on each other are sufficiently equal. Additionally, because they have to be equal, they have to be low since they are determined by the people who impose the lowest levels of risk upon others (the 'risk-dictators'). However, it is not easy to determine these levels since people not only directly impose risk on each other but also are implicated in a whole range of indirect risk-imposition associated with societal goods and services.

Notwithstanding these difficulties, we can still draw three implications from the contractualist perspective on acceptable risk imposition that will be valuable input for the current debate. First, the equality requirement entails a strong presumption in favour of risk-reduction. Using my central example of the CEO and the nurse, I have shown that adherence to the separateness of persons and a commitment not to justify risks based on the interpersonal aggregation of benefits has far-reaching consequences. Applied to our current society, it involves a drastic call to try much harder to take precautions and reduce risks. Objections that this would be too costly and would have a negative cost-to-benefit ratio miss the mark. In the next chapter, we will see that in the domain of mobility, many cost-effective measures to reduce the toll of traffic in terms of casualties and injuries still have not been implemented. More importantly, appealing to cost-benefit analysis (CBA) misses the entire point of this chapter, which is that contractualism aims to avoid the interpersonal aggregation that is centre stage to cost-benefit analysis.[91]

As a second implication, Scanlon's appeal to a positive intra-personal aggregation of agency benefits and risk burdens over the course of one's life is instructive. Today, the question of acceptable risk imposition is often posed with respect to single actions, projects, or practices. However, given that the moral acceptability of mutual risk imposition depends on the cumulative risk burden versus the cumulative agency benefits along one's entire life, there must be a shift in focus towards the sum of risk that (groups of similar) individuals incur from different sources. Consequently, it might be that the cumulative effect of different risk-generating practices—that in isolation would seem morally permissible—is unacceptable. It is not immediately clear what that would mean for those individual practices, but this shift in focus should have consequences for our current procedures for regulating risk. At first sight, it seems likely that this would be an improvement for vulnerable people (such as those suffering from heart and lung diseases) who bear the heaviest burden from the accumulation of risks from different sources.

Let me draw a final implication. Ultimately, Scanlon's contractualism seems to be committed to including all citizens of the earth in a global system of equal mutual risk imposition. Given that the consequences of our modern way of life, including risks, extend around the globe, we must be able to justify them to each world citizen since each has equal moral status.

Since, if it should be the case that no one has reason to object to principles regulating risk imposition, moral theories committed to each person's equal moral status should exclude no one. A well-known example with regard to contractualism concerns the case of an Amish farmer, who does not benefit from air travel over his land. According to Ashford, many seem to have the intuition that air -travel is morally acceptable, "even if a few persons are randomly killed by it, who could not have expected to benefit from it, provided there are significant, even though smaller, benefits to a huge number of others"(Ashford: 2003, 301). Instead of saving this intuition by appeal to (allegedly inevitable) inter-personal aggregation, Scanlonian contractualism requires us to investigate whether air travel can be or become part of a system of equal mutual

[91] For further discussion, see (Keating: 2003; Lenman: 2008)

risk imposition that *includes all* potential victims. If we cannot do so, we have nothing to say to those who simply refuse to be potential victims for the benefits of air-travellers.[92] This inclusion boils down to making makes it the case and showing that risk-regulating principles, insofar they apply to actions that have global consequences, are intra-personally beneficial to each world citizen. To succeed, appeal to the benefits of indirect risk-impositions allowed by these principles most likely will do a great deal of the justificatory work.

These three implications also are clearly relevant for managing risk in connection to persuasive technology. They are sufficiently general to apply to risks generated by the use of persuasive technology. In addition, they help determine whether some type or risk or risk-generating practice is morally acceptable or whether citizens can be obligated to use persuasive technology as a means of risk reduction. At the same time, in its current state, contractualism is too indeterminate to enable a full and well-founded answer to concrete issues in applied ethics. As discussed in this chapter, this indeterminacy has two main sources. First, there is the issue of whether or not agency costs are to be included in contractualist reasoning. Second, in our society, a large share of risk-imposition is indirect, which makes it difficult to determine the levels of acceptable mutual risk imposition.

For these reasons, we need additional considerations for guiding moral deliberation on practical issues. The following four considerations are prominent in the literature (Cranor: 2007; Hansson: 2003; Hayenhjelm: 2012; MacLean: 1986; Railton: 1985). The first concerns the nature and magnitude of the goods and benefits that result from risk-imposing action, compared to the nature and magnitude of the risks. The risks must be worthwhile to take. Second, it is very important how these benefits and burdens are distributed over the various individuals involved. This distribution must be fair in ways that need to be specified. Third, if those upon whom a risk is imposed give their consent, whether actually, tacitly, or perhaps even merely hypothetically, this consent adds significantly to the moral acceptability of the risk in question. Fourth, as

[92] But perhaps, as Philip Nickel suggested to me, we could appeal to their moral motivation. Scanlon's contractualism presupposes morally motivated participants, who desire to justify themselves to each concerned. Part of such moral motivation may be charity in accepting small risks in order to allow others freedom to pursue their projects. Perhaps that Amish farmers might be willing to accept the risk of planes falling on their grounds, but likely not the risks of, for example, climate change and environmental pollution.

also explained in this chapter, taking due precaution is a further method of enhancing this acceptability.

It is difficult to specify in advance a precise method according to which this plurality of considerations can best guide our judgments of acceptable risks. However, once applied to a concrete moral or political issue, they often give remarkably clear guidance. The next chapter attempts to show this for the case of Intelligent Speed Adaptation as a means of reducing driving risks.

7 The moral case for Intelligent Speed Adaptation[93]

7.1. Introduction

In this chapter, I will argue that there is a moral case for setting mandatory speed alerts and speed limiters in all cars. These technologies are fairly intrusive. Nevertheless, my claim is that we should accept these measures in our cars to solve a major problem in road safety: speeding. In 2010, in Europe, more than 30,000 people were killed and 1.4 million were injured in road traffic, with speeding as a major cause. Current enforcement measures work to some extent but are clearly not sufficient. Intelligent Speed Adaptation (ISA) systems are highly effective additional measures to counter speeding. Advisory ISA warns drivers if they transgress the speed limits. Limiting ISA makes speeding impossible, and consequently this technology can prevent up to 50 % of fatal accidents.[94]

Intelligent Speed Adaptation is indispensable for reducing the risks of car driving to a more acceptable level. Many philosophers uncritically refer to driving as an example of acceptable risk imposition (Hansson: 2003; Hayenhjelm & Wolff: 2012; McCarthy: 1997). The benefits of car driving are considered to justify the risks involved, which are perceived as being relatively low. Car driving is regarded as a morally acceptable practice from which we all benefit. However, as I will argue below, this view is problematic even with regard to lawful car driving. Moreover, in appealing to car driving as an example of acceptable risk imposition, one fails to appreciate the fact that the practice involves massive transgressions of the rules. Pedestrians, cyclists, and lawful drivers have good reason to reject the risks involved in our *actual* car driving practice. No tacit consent to the risks of driving can be inferred from individuals' choice to walk, cycle, and drive (cf. Thomson, 1985a).

[93] This chapter has been accepted for publication in the *Journal of Applied Philosophy*

[94] See section 7.2 and 7.3 for references.

Speeding thus imposes excess risk of harm to life and limb. Speeding involves transgressions of democratically accepted rules for traffic risk regulation, which ideally represent a fair weighing of driving risks and benefits. Thus, the resulting harm would be wrongdoing. Therefore, it appears that, as in the criminalisation of speeding, ISA is justified by the prevention of wrongful harm to others. Insofar as ISA is an effective criminal enforcement method, there appears to be a straightforward moral case for ISA.

However, several objections to ISA are raised. In particular limiting ISA, which makes speeding technically impossible, meets strong resistance from many drivers (Vlassenroot, Marchau, De Mol, Brookhuis, & Witlox: 2011). In the popular press, readers raise the concern that ISA introduces new safety risks that they are unwilling to accept.[95] Additionally, they claim that ISA is an intolerable interference with drivers' liberty. Legal scholars have argued that the enforcement of the law by means of technologies such as ISA goes beyond the law itself, because this type of enforcement makes civil disobedience impossible and leaves no room for interpretation of the law. People obey the law because they are unable to do otherwise and not as an exercise of moral agency. Consequently, ISA-like technologies might endanger the development of the moral capacities of citizens and reduce their human dignity (Brownsword: 2005; Yeung: 2011). Even worse, arguments similar to those that justify ISA appear to justify an entire range of other behaviour-regulating technologies.

In addition to its direct practical relevance, then, ISA is an interesting example of a broader trend in regulation, namely so-called 'techno-regulation', or 'design-based regulation'. Techno-regulation refers to regulating and channelling conduct "by relying on [...] integrated technology or design" (Brownsword: 2005, 2). Techno-regulation that tries to influence people while leaving them free choice can be classified as a so-called persuasive technology (Fogg: 2003). Advisory ISA, which warns drivers if they transgress the speed limits, falls within this category. Techno-regulation that makes undesirable behaviour impossible or a desired behaviour the only option could be called a limiting or forcing technology. Limiting ISA, which makes speeding impossible, is a clear example of such a technology.[96]

[95] See, for example, reader responses to an article in the Mail On Sunday (Owen: 2013). In fact there was no EU plan for mandatory ISA that the article criticizes to oppose

[96] In the literature one can also find other terms for these versions of ISA, e.g. 'mandatory ISA' instead of limiting ISA, and 'voluntary ISA' instead of advisory ISA.

This chapter is organised as follows. In the next section, I will argue that car driving poses morally problematic risks. In section 7.3, I will explain how ISA can reduce the risks of driving by reducing the incidence of speeding. In section 7.4, I will argue that harm prevention justifies ISA in much the same way as the criminalisation of speeding. In section 7.5, I argue that drivers are morally required to accept the additional safety risks introduced by ISA to make driving safer for others. In section 7.6, I will extensively evaluate the concern that ISA erodes the moral agency and responsibility of drivers. In section 7.7, I argue that the reasons for implementing ISA do not commit us to 'designing out' all forms of undesirable behaviour. ISA is not a first step to the virtual disappearance of citizens' liberty. Finally, I conclude that there is a moral case for all of the different versions of ISA (7.8). Amongst these versions, limiting ISA with a highly restricted option to override the system appears to be the most favourable.

7.2. Driving risks are morally problematic

Although many of us drive, closer inspection shows that current car driving is morally problematic. Driving risks may *appear* to be acceptable because these risks are the result of a (democratic) agreement about risks and benefits that is viewed as working for the benefit of all. However, this impression is false. Pedestrians, cyclists, and some drivers, especially lawful drivers, have at least three reasons to reject the driving risks that are imposed on them.[97] First, the risks are, in fact, substantial *and* higher than necessary to receive benefits. This is most clearly problematic from a consequentialist perspective. In addition, drivers impose highly non-reciprocal risks on pedestrians, cyclists, and some other drivers. The distribution of benefits and burdens is unfair. Finally, precautionary measures, such as airbags, are distributed unevenly amongst different road users. These latter two reasons have to do with considerations of fairness that are central to a deontological perspective.

ISA is relevant to each of these three problematic characteristics of car driving practices as follows. Speeding contributes strongly to unnecessarily high risk levels. Furthermore, speeding severely aggravates the consequences of both the

[97] Here I focus on safety risks, and leave out environmental and health risks, which concern other stakeholders as well.

non-reciprocal character of driving risks and the absence of due precaution. I will elaborate on each characteristic in turn.

First, the statistics make clear that driving risks are indeed substantial: road traffic is a major cause of death and injury all over the world. Although the risks imposed by a single car trip may appear to be low, it should be noted that driving is not an isolated action. Driving is a regular practice for which the *cumulative risks* imposed on road users are the most appropriate subject for moral evaluation (Husak: 2004). In 2010, approximately 30,400 European Union citizens were killed in road traffic accidents, of which approximately 19 % were pedestrians and 7 % were cyclists (European Road Safety Observatory: 2012b). In addition, approximately 1.4 million people were reported as injured (European Road Safety Observatory: 2012a). In 2011, an estimated 250,000 people were seriously injured, of which a significant number became permanently disabled.[98] The estimated economic cost of these accidents is 2 % of the European yearly GDP.[99] Husak, writing on this problem in the context of the US, where it is worse than in Europe, rightly remarks that driving is the riskiest activity in which the vast majority of Americans routinely engage. It is safe to predict that if the typical reader of these pages (directly) kills or seriously injures another person, his weapon is likely to be a motor vehicle. (Husak: 2004, 355)

A quick glance at the data provided by the World Health Organization shows that the figures in the rest of the world are even worse than in the US.[100] Driving poses significant risks to human health and life.[101]

[98] http://ec.europa.eu/transport/road_safety/topics/serious_injuries/index_en.htm (Last accessed 11-12-2014). Real figures of (serious) injuries may be different and are most likely larger, because different reporting criteria and different definitions of serious injury are used across the EU. So, *estimates*, based on the European Injury Database, indicate that "more than four million people are injured annually in road traffic accidents in Europe, one million of whom have to be admitted to hospital" (European Road Safety Observatory: 2012a).

[99] http://ec.europa.eu/transport/road_safety/topics/serious_injuries/index_en.htm (Last accessed 11-12-2014).

[100] See http://www.who.int/mediacentre/factsheets/fs358/en/index.html (Last accessed 11-12-2014).

[101] *Contra* McCarthy, who writes, without providing arguments, that "Driving by someone and thereby imposing, say, a one in ten million risk of death on her seems to be an action of little moral significance" (McCarthy: 1997, p. 211). As said, we should not focus on a single drive, but on the practice as a whole. In 2010, 0.64 % of EU deaths were traffic deaths (European Road Safety Observatory: 2012a). Thus, in 2010, the chance to die in traffic extrapolated to the

The massive transgression of traffic rules is one of the causes of these unnecessarily high risk levels. In most countries, between 20 and 40 % of cars exceed the speed limit, depending on the road type. Speeding is even more common in other countries Speeding is an *avoidable* major contributing cause of accidents (OECD & European Conference of Ministers of Transport: 2006, 55).

Philosophers who present car driving as an example of acceptable risk imposition may respond to my argument so far by emphasizing that they are referring to lawful driving. Suppose that lawful car driving does indeed involve a level of risk that is morally acceptable. Then, their reply merely serves to accentuate the need to bring actual driving practice much closer to lawful driving. In fact, the huge allowance for unlawful driving constitutes the highest risk of car driving and is a good reason to reject current driving practices.

The second reason to reject driving risks applies to both unlawful and lawful driving: the risks are highly non-reciprocal. Risks and benefits are not distributed fairly among different road users. This is most obviously the case for risks imposed by drivers on pedestrians and cyclists. However, the risks of crashes can be distributed highly unevenly also among drivers because different cars have different masses and can thus be ‘crash-incompatible’.[102]

This non-reciprocal character of driving risks cannot be justified by the distribution of benefits. On the contrary, the resulting benefits of driving are primarily derived by the parties with the lowest risk, the drivers. The numerous pedestrians who do not drive receive no benefits from a significant part of all trips made by car, not even indirect benefits. Drivers of heavy cars receive safety benefits at the expense of drivers of lighter cars. Crucially, these risk impositions need not be as non-reciprocal as they currently are for car driving to deliver its benefits. For, cars need not be heavy to travel from A to B. And, much more can be done to protect vulnerable road users. In the EU, pedestrian and cyclist safety have only recently become important design considerations for car manufacturers, spurred by 2007 EU legislation that has (finally) addressed the problem of

lifetime of a EU citizen was 1 in 156. I suspect that many people would not accept this risk if they were aware of its magnitude.

[102] If the mass difference increases, the lethal risk for the driver of the lighter car increases, while the risk for the driver of the heavier car decreases proportionally. A driver of a car that weighs 1079 kg, crashing with a 2100 kg car, has 21 times more chance of dying than the driver of the 2100 kg ca. In addition to the distribution problem, the cumulative risks increase as well. For the Netherlands, it is estimated that 25 % of all deaths resulting from crashes between passenger vehicles are caused by mass differences (Berends: 2009).

vulnerable road users. Therefore, drivers impose many partly avoidable risks on other road users, while receiving the lion's share of the benefits.

The third reason not to accept car driving risks relates to the above discussion of the non-reciprocity of driving risk and regards the uneven distribution of precautionary measures amongst road users. Precaution is a crucial condition for acceptable risk imposition: have those imposing the risk taken due precautions directed towards *each* of the affected persons?[103] Unfortunately, the safety measures with which cars are equipped are one-sided in being designed for drivers and passengers. For example, although airbags for drivers have been more or less standard for approximately two decades, there are virtually no cars with airbags for pedestrians. Car driving is, at least in Europe, changing for the better but, in its current form, it falls short of taking reasonable precaution for other road users. Drivers cannot make the argument that they are not responsible for car design. Manufacturers are highly sensitive to consumer wishes that these are willing to pay for. If enough drivers were genuinely worried about pedestrian safety, car manufacturers would respond.

To conclude, although it is certainly morally significant whether or not drivers obey the law, even lawful driving is morally problematic. Pedestrians, cyclists, and some car-drivers have more than sufficient reason to reject current driving risks.[104] As already noted, the mere fact that they use the road cannot be taken as (implicit) consent to these risks. Often they have no alternative. Furthermore, they may be largely unaware of several of the facts of driving risks as discussed in this section.

The aforementioned risks reinforce each other. Although speeding is risky in itself, crash incompatibility and a lack of precautionary equipment increase the risks of speeding enormously. This mutual reinforcement amplifies the moral urgency of reducing the incidence of speeding, thus strengthening the case for

[103] For an elaboration, see Scanlon (2000, 208–209) and Lenman (2008).

[104] See also Husak (2004), who argues, that considerable part of our driving practice, even where in accordance to the law, should be regarded as morally wrong. He argues from the facts that many trips are 'frivolous', i.e. are not necessary, and are taken in highly crash-incompatible vehicles such as SUVs. (Husak could have added that all cars are crash-incompatible with cyclists and pedestrians). In addition, he explains why his claim may appear so counterintuitive to many of us. People underestimate risks they believe to control, "people downplay the risks of conduct they hold to be beneficial", and "people tend to regard risky conduct permissible if they engage in it themselves" (p. 368). It seems that philosophers referring to car-driving as an example of acceptable risk impositions felt prey to some of these biases.

ISA. In the next section, I will explain how ISA can reduce driving risks significantly.

7.3. Significant risk reduction through Intelligent Speed Adaptation

ISA leads to a major reduction of driving risk by reducing the incidence of speeding. The effectiveness of ISA can be explained in terms of the strong and empirically well-established relationship between speed and accident risk. Higher speed increases the risk of accidents by increasing the stopping distance and decreasing manoeuvrability and the available reaction time. In addition, higher speed always increases the severity of accidents, thus increasing the likelihood that an accident will result in injury or fatality. Consequently, a tiny reduction of the mean speed from 120 km/h to 119 km/h already results in 3.8 % fewer fatalities. In addition to a higher mean speed, also a higher variance in the speed of different cars increases accident risks (Aarts & van Schagen: 2006; Elvik: 2009: 2014).

There are several variants of the ISA. All of these variants employ technology, most often global positioning technology, that locates a vehicle. The location of the vehicle is coupled to a database that provides the corresponding speed limit, which enables feedback to be provided to the driver. Advisory variants of ISA display the speed limit and warn the driver if he exceeds the limit. Supportive ISA limits the speed (e.g., via the engine management system or a gas pedal that exerts upward pressure) but can be overridden at any time. Limiting versions operate in comparable ways but go beyond supportive ISA by limiting the driving speed *without* allowing the driver to override the system. Interestingly, ISA technologies enable governments to go beyond the system of fixed speed limits, and to work with dynamic speed limits. Dynamic speed limits can vary with the time of the day, the weather, traffic load, and other conditions.

Extensive research predicts sizable absolute risk reductions upon implementing ISA. Numerous driving simulator experiments and field tests with ISA have consistently shown the following effects: a decrease in the mean speed, a decrease in the speed differences among cars, and a decrease in transgressions of the speed limit (Lai, Carsten, & Tate: 2012; Vlassenroot et al.: 2007). It is predicted that the use of advisory, supportive, and limiting ISA in all cars would prevent 2.7, 12.0, and 28.9 %, respectively, of injury accidents in the UK in which a car is involved (Lai et al.: 2012). This 28.9 % accident reduction for limiting ISA can be extrapolated to a 50 % reduction in fatal accidents (Carsten:

2012).[105] An Australian trial with an advisory ISA predicted a reduction of 5.9 % for injuries and 8.4 % for fatal accidents (NSW Centre for Road Safety: 2010, 98). Calculations that were based on data from a Dutch trial with limiting ISA predicted a 25-30 % reduction of fatal accidents (Oei & Polak: 2002). These estimates are calculated on the basis of real-world speed data from ISA field-trials, empirical quantitative models of the relations between speed and accidents (as referred to above), and accident statistics for the respective countries.[106] In a traffic system with dynamic speed limits the reductions will most likely be even larger (Carsten & Tate: 2005). Data giving predicted absolute numbers of lives saved are absent, but some rough estimations can be made. Given the approximately 30.000 yearly road traffic fatalities in Europe, limiting ISA will save thousands of lives. For individual countries this will typically be a few hundreds.

To appreciate the precise scope of my claims, it is important to note that the speed at which ISA systems intervene, either by warning or limiting speed, is a matter of choice. This speed could be the legal maximum speed or a higher speed.[107] In most ISA trials, the intervention speed was the legal speed limit, sometimes with a very small margin. In this chapter, all my arguments apply to ISA that intervenes at the legal speed limit because this is the strongest and thus most interesting case for which to argue. In addition, this version of ISA matches the current practices of speed limit enforcement in several countries. Legal speed limits have a particular significance regarding a democratically accepted level of risk. Nonetheless, readers who are ultimately unconvinced by my arguments are invited to consider how their ethical evaluation of ISA would change for an increased speed level at which ISA intervenes.

[105] Comparable extrapolations will apply to the other versions, but are not provided by Carsten.

[106] It should be noted that different studies use slightly different though comparable ISA systems, and somewhat different methodologies, and apply to different countries, for which the magnitude and characteristics of the speeding problem may differ. Still, the results of those studies constitute a reliable indication of the effectiveness of the different versions of ISA. Unfortunately, there is no study that has both injury and fatal accident savings based on field trial data for all three versions of ISA.

[107] I thank Philip Nickel for emphasizing this point.

7.4. Harm prevention justifies ISA

Obligatory ISA is justified by the aim of preventing harm in much the same way as the current criminalisation of speeding. In many countries, laws have been passed that set speed limits and criminalise the transgression of these limits. The justification of these laws follows relatively straightforwardly from the fact that they prevent severe harm. According to Feinberg's well-known formulation of this idea that the prevention of harm to others supports criminalisation, "it is always a good reason in support of penal legislation that it would probably be effective in preventing (eliminating, reducing) harm to persons other than the actor (the one prohibited from acting) *and* there is probably no other means that is equally effective at no greater cost to other values" (Feinberg: 1987, 26). This rationale for criminalisation includes the aim of preventing significant *risk* of harm, since many harms are the result of accidental events. Harm prevention is only *a* reason in support of criminalisation, and not a sufficient or even a necessary reason, for we do not criminalise all behaviours that causes harm or risk of harm to others.

Speed limits that are rightly set play an import role in determining which driving behaviours involve such risks of harm to other road users that it ought to be criminalised. Driving below the speed limit creates the very same risk of accidents as speeding, only of a lesser magnitude. Nevertheless, we do not prohibit all driving, because flexible and time-efficient mobility has high value. Setting the speed limits, then, involves a judgment based on the social benefits of driving at a certain speed and the distribution of these benefits, and the safety risks and their distribution among different road users. Thus, we weigh driving risks against the benefits and also consider whether risks and benefits are distributed in a sufficiently fair way.

Given morally justified speed limits, exceeding them *wrongfully* imposes *excess* risk of accidents. Again, these excess risk are substantial: consider, for example, that an increase in the mean speed at highways from 63 to 70 mph leads to 62 % more fatalities,[108] together with the high incidence of speeding, between 20 % and 40 % for most countries (OECD & European Conference of Ministers of Transport: 2006). These substantial additional risks cannot be justified, because the additional benefits no longer outweigh them, and also the distribution of risks and benefits becomes too problematic. Our right not to be

[108] Calculated on the basis of (Elvik: 2014).

harmed bodily or killed is (one of) the most widely acknowledged and most important of our rights. The interest of bodily integrity, protected by this right, is a profoundly basic interest that all humans share and on which nearly all of their other interests crucially depend. Therefore, the interest of traffic safety for all participants easily outweighs the drivers' interests that are associated with speeding, such as saving travel time or the enjoyment of driving fast. Thus, speeding is a serious wrong. It imposes unjustified substantial risk of grave harm to others,[109] which is sufficient reason to criminalise speeding. At the very least, the justification of the relevant road traffic laws that many countries have adopted can be understood along the lines sketched here.

It might be objected that some instances of speeding are not of a nature that they should be considered as serious wrongs. For example, "am I really imposing a substantially increased risk of harm to others when I drive a few mph over the speed limit for 20 miles along a deserted highway under excellent driving conditions?".[110] The following responses are possible. First, as I argued above driving at the legal speed limits *already* involves substantial risk, which however, according to our collective judgment, is justified by the mobility benefits. Therefore, the *increase* of this risk need not be substantial for speeding to be wrong. Second, even if it were conceded that some incidences of speeding do not constitute a serious wrong, or a wrong at all, speed limits are still justified by the prevention of harm to others on the level of the driving practice as a whole. No proof that every single case of speeding is morally wrong is needed to justify

[109] Of course, not all the fatalities and injuries mentioned in section 7.2 qualify as harm to others. Exact data appear absent here. Approximately one third of all fatalities in the EU during the period 2001-2010 were caused by "single vehicle accidents" (see European Road Safety Observatory: 2012c). For the remaining accidents, part of the casualties and injuries regard drivers who caused the accident, and thus harm themselves. Note however that passenger harm classifies as harm to others in case it results from speeding or other non-lawful driving. In countries where cycling is common, the category 'injured' involves a significant number of accidents in which no car driver is involved, but a cyclist. Still the number of fatalities and injuries that qualify as "harm to others" will be large enough to justify my argument for ISA on the basis of the prevention of harm to others alone. If we are willing to also accept the prevention of "harm to self" as a justifying ground for ISA, then of course all fatalities and injuries count. Although in this chapter I will not argue for ISA on paternalistic grounds, I do think that this would be plausible in the case of children, and young adults, which are involved in a disproportionally large part of all accidents.

[110] This is how an anonymous reviewer put it. I thank the reviewer for pressing this objection to me.

uniform speed limits, and the same is the case for the justification of ISA. Third, ISA enables dynamic speed limits that could be set higher under the conditions mentioned in the objection.

At this point, it is useful to distinguish between a weaker and a stronger claim defended in this chapter. The weaker, conditional, claim is that, *if* current levels of the speed limits are justified (representing a just weighing of risks and benefits), then obligatory ISA is justified by the harm it prevents.[111] The stronger claim is that obligatory ISA is in fact justified in this way. The reason is that, by and large, current speed limits are too high, rather than too low, such that if the limits are exceeded, this clearly creates a risk of harm that ought to be prevented. As discussed in section 7.2 above, driving risks are substantial, the protection of vulnerable road users is far less than reasonable and feasible, while the benefits of driving go one-sidedly to drivers who create the risks. Many cost-effective road safety measures have been identified that still wait for implementation (SafetyNet: 2009). As long as these problematic aspects of current driving risks are not satisfactorily addressed, our speed limits are most likely not too low. A discussion as to how the speed limits should be set exactly is, however, far beyond the scope of this chapter. This would have to include an extensive weighing of cost and benefits, as well as a determination of the relative importance of fairness considerations. For present purposes, this discussion of the justification of speed limits should suffice.

It is reasonable to hold that harm prevention does not only support criminalisation, but techno-regulation, such as ISA, as well. This is in the spirit of the broader Millian liberal core idea that the only legitimate reason for which the state may interfere with the liberty of its citizens is to prevent harm to others (Mill: 1985, 68). Techno-regulation is another way in which the state can interfere with the liberty of citizens to prevent harm to others. Like criminalisation, techno-regulation can often constitute a strong restriction of citizens' liberty, which, in principle, should only be imposed by the state.

However, some may doubt that preventing harm to other road users justifies a policy of making ISA mandatory in all cars. It could be argued that because physical limitations of liberty extend beyond legal limitations, a stronger justification is needed for the former. This view is hard to defend as a general claim,

[111] I thank two anonymous reviewers for emphasizing the central importance of correct speed limits for the justification of ISA.

because there are many intentionally designed physical limitations on how we can act that are completely non-problematic. In particular our built environment contains many such limitations. For example, elevated pavements exclude drivers from space dedicated to pedestrians, and iron fences strongly inhibit unauthorized climbing of pylons.[112]

Nevertheless, I will address three respects in which ISA could be thought to require more justification than criminalising speeding First, the outcome of balancing the interests of different road users is not fundamentally altered by implementing ISA in addition to the criminalisation of speeding. Road traffic law enforcement is a heavy burden on the criminal justice system. Imprisoning offenders incurs serious public costs, such as the stigmatisation of wrongdoers and financial costs (Fahlquist: 2009), that would be reduced by the implementation of ISA. However, driving will become less pleasurable for a significant portion of drivers who may experience the warning signals from advisory ISA as annoying and intrusive or because of the impossibility of speeding with limiting ISA. These drivers may consider the reduction in their enjoyment of the driving experience as a significant loss, and any government that aims to successfully implement ISA should address this consideration. At the same time, research has shown considerable variation in drivers' attitudes towards ISA (Vlassenroot et al.: 2011). A different group of drivers may feel supported by this system and find driving much more pleasurable knowing that all cars have limiting ISA.

It is of crucial importance to appreciate that ISA does not affect the central interests of drivers. Drivers can still travel to their destination in a flexible, time-efficient, and private manner, which enhances personal autonomy (cf. Lomasky: 1997). ISA systems do not interfere with this opportunity; they only make it safer for all of the parties involved. Drivers who experience a significant loss of freedom and pleasure from the use of ISA could merely be individuals who never took the legal limits seriously in the first place. Feinberg's analysis of the 'fecundity' of liberties applies here (Feinberg: 1987, 206–214). Mobility as such

[112] The need for justification of legal and technological (physical) limits to driving speed arises merely because these constrain an option that is *perceived* as significant by many people. But one could still raise the question as to whether that perception might be misguided, or highly contingent. Imagine that we were to design cars from scratch and that no one had prior experience with car-driving. If these cars *by design* could not exceed the speed limit, I doubt whether that would be perceived as a serious liberty limitation.

is an option that leads to many other valuable options, whereas speeding primarily leads to enjoyment and small travel-time savings.

However, speeding may save lives in cases of emergency. Given that in most cases calling ambulance services is the best option, the number of lives saved by limiting ISA will significantly outweigh the number of lives saved by emergency speeding. Furthermore, it is not at all clear that a need to speed in cases of emergency would trump others' right to protection against speeding risks. Still, limiting ISA would rule out some rare cases in which speeding, within certain plausible limits, appears to be justified. To conclude, in case of both the criminalisation of speeding and ISA, the safety interests of all road users outweigh the interests of the drivers in speeding.

The second point of comparison between the justification of criminalisation and ISA regards their effectiveness. As noted above, criminalisation can only be justified if it is effective in preventing harm. This implies that greater effectiveness in preventing harm to others provides stronger support for limiting liberty. Limiting ISA could prevent up to 30-50 % of fatal accidents (section 7.3), enough of which are not the victim's own fault and thus qualify as harm to others. At least in the particular case of ISA, stronger support for liberty limitation can plausibly also be interpreted as support for greater liberty limitation. This interpretation seems plausible because even though a physical speed limit goes indeed beyond a legal limit, the liberty to speed is of minor fecundity.

Finally, ISA may come with "greater cost to other values" (Feinberg: 1987, 26) than mere criminalisation of speeding. In particular, ISA may threaten legislative values. One concern regarding techno-regulation in general is that such regulation could negatively affect the ideal of 'legality'. Legality refers to the concept that "law should be viewed as the product of an interplay of purposive orientations between the citizen and his government [and not] as a one-way projection of authority, originating with government and imposing itself upon the citizen".[113]

Legality poses a potential problem for ISA because it is often assumed, including by politicians (Carsten: 2012), that ISA lacks societal support. If this assumption is correct, the policy of making ISA mandatory for all drivers would fall short of the moral ideal that citizens express their autonomy via democratic lawmaking. However, advisory ISA is supported by a majority of drivers,

[113] Lon Fuller, *The Morality of Law*, cited in (Brownsword & Goodwin: 2012, 451).

whereas a significant minority supports stronger (i.e., supportive and limiting) versions of ISA (NSW Centre for Road Safety: 2010; Vlassenroot et al.: 2011). Moreover, what counts is the view of *all* citizens, not only the view of drivers. It seems not unlikely that the overall societal support for limiting ISA is greater than that by drivers alone, though I know of no data here.

Other important legislative values, such as 'transparency' and 'accountability' (Brownsword & Goodwin: 2012), can also be safeguarded. The decision to make any version of ISA mandatory should involve democratic procedures in the same way as the underlying traffic laws and legal speed limits.

To conclude, ISA can be justified on the basis of preventing harm to other road users, similar to the criminalisation of speeding. As each of the objections to ISA that were mentioned in the introduction hinges on certain "cost to other values" arising from ISA, I will now discuss these objections in turn.

7.5. Objection I: ISA introduces new safety risks

One objection against ISA that is frequently voiced in the popular press is that this technology will introduce new safety risks from sources such as technological malfunctioning, negative adaptation of driving behaviour to ISA (e.g., tailgating, higher degrees of frustration), and increased time needed for overtaking (Jamson, Chorlton, & Carsten: 2012; van der Pas, Marchau, Walker, van Wee, & Vlassenroot: 2012). One could imagine what might happen if, for example, the localisation technology of limiting ISA were to erroneously locate a car along a stretch of a road with a speed limit of 50 km/h, when the driver is in fact driving on an adjacent, parallel highway. The objection grants drivers the right not to accept such additional and partially involuntary risks and thus to reject ISA. This objection is, however, not valid.

The safety risks that accompany ISA are real, although their magnitude is uncertain, and depends on several factors. The extent of behavioural adaptation, for example, depends on the scale of implementation: is the driver one of a few driving with ISA or is ISA the standard? Overtaking may become more dangerous with ISA, but drivers are fully in control of deciding whether to overtake. If ISA leads to less overtaking, safety may, in fact, increase. No sound estimates of magnitudes are currently available for many of these risks; however, in general, the level of uncertainty, as judged by experts, is higher for limiting ISA than advisory ISA (van der Pas et al.: 2012). Two ways of gaining more knowledge and addressing these risks are to do more research and begin monitored implementation, most likely with advisory versions of ISA (van der Pas et al.: 2012).

Despite the uncertainty in the aforementioned risks, the numerous trials that have been conducted since the 1990s do not provide any evidence that the magnitude of these safety risks would be considerable. It is therefore highly plausible to assume that the large gains in safety that result from decreased speeding will significantly outweigh any losses from ISA safety risks. Although it is important to validate this claim during ISA implementation, in the remainder of this section I will assume that it holds.

Accordingly, from a consequentialist perspective on the ethics of risk, drivers cannot refuse to accept safety risks that arise from ISA on moral grounds, because ISA can be expected to produce a significant overall increase in safety. Furthermore, this expectation also applies to drivers individually. Thus, drivers cannot protest that their well-being is sacrificed for the sake of the well-being of other road users. Nevertheless, they could maintain that they prefer the higher risks that arise from others' speeding above the risk of, for example, malfunctioning ISA technology. These drivers may strongly dislike the fact that they cannot control that risk. As Teuber argues, we are not merely concerned with the level of the risks we bear, but, as autonomous persons, we also value control over these risks (Teuber: 1990). Therefore, from a deontological perspective, it is crucial that people consent to the risks imposed on them (be it perhaps only hypothetically).

However, from this perspective it also follows that a driver cannot reasonably refuse to accept additional safety risks that arise from ISA. In section 7.3, we saw that drivers impose substantial, avoidable, and non-reciprocal risks on vulnerable road users without taking reasonable precautionary measures. Again, reasonable precaution is an important condition for acceptable risk imposition. ISA is such a precautionary measure that is highly effective. In fact, ISA is indispensable for reducing the risks of driving practices down to a level that is acceptable to all road users. Therefore, drivers should accept the imposition of some (likely small) risks on *themselves* to reduce the substantial risk they impose on *others*.

Nonetheless, a lawful driver might object to this line of reasoning by arguing that she never speeds or that advisory ISA would be sufficient to prevent her from speeding. She is certainly right, but the relevant question is whether her reasons to reject ISA are stronger than the reasons for which an individual pedestrian or cyclist could reject the alternative, i.e., cars not equipped with ISA. From my discussions of driving risks (section 7.2) and ISA risks, it should be clear that the vulnerable road users have the strongest case. The only way for the lawful driver to have the practice of driving, is to accept measures that prevent fellow drivers from speeding. To conclude, the safety risk objection to ISA does not hold.

7.6. Objection II: ISA erodes the moral agency and responsibility of drivers

Roger Brownsword is an eloquent spokesman for the concern that techno-regulation, such as ISA, makes disobedience impossible and thus reduces and erodes citizens' moral agency and responsibility. Brownsword discusses the example of a fully automatic car and concludes that its implementation would mean a reduction of opportunities "for choosing agents to respect the vulnerability of others". He argues that such a car is the first step towards the "corrosion of the moral community". This moral community essentially presupposes that people are vulnerable rights-holders whose legitimate interests can be harmed. However, these people are also duty-bearers who "have at least some opportunity to inflict harm on right-holders" (Brownsword: 2005, 19). I will label these opportunities for inflicting harm 'opportunities for moral agency' because what we value positively is that agents understand and act on the basis of moral reasons. These opportunities for moral agency presuppose that agents can act otherwise and cause harm.

How should ISA be evaluated in light of this concern? I will show that although ISA does take moral agency and responsibility away from drivers, this is justified by the safety benefits ISA provides. Karen Yeung analyses the effect of hypothetical automatic braking technology, which is activated by red traffic lights, on drivers' moral agency and responsibility (Yeung: 2011). I will draw from her example to analyse ISA systems. Let us consider a version of advisory ISA that emits a clear warning signal for 5 seconds for every transgression of the speed limit. How does this ISA affect the moral agency of the following three different stylised types of drivers: the vicious driver who only acts on prudential reasons, the ordinary driver who typically acts on a mix of prudential and moral reasons, and the virtuous driver who always acts on moral reasons?[114]

[114] Adapted from Yeung (op. cit., 11-15) by application of the fourfold typology of agents in Brownsword and Goodwin (op. cit. 437–438), leaving out the fourth.

Advisory ISA will elicit whichever reasons for not speeding are most important and accessible for a certain driver. The vicious and ordinary drivers are reminded of their self-interested reasons for not speeding, such as not getting fined and not getting hurt in an accident. The ordinary and the virtuous driver are, in situations in which they are not sufficiently attentive, reminded of their moral reasons for not speeding: having regard for the safety and well-being of others. As long as advisory ISA is not equipped with a detection function, the technology will not emphasise prudential reasons at the cost of moral reasons (cf. Brownsword & Goodwin: 2012, 436–439). In addition, ISA would prevent each driver from speeding unintentionally, thus benefitting each (cf. Yeung: 2011, 14). We see that the opportunity for exercising moral agency is not reduced for any of these three drivers but that the driver is in fact supported in acting morally. Somewhat surprisingly, advisory ISA is a piece of techno-regulation that *promotes* rather than erodes the flourishing of the moral community.

The picture is different for limiting ISA. The decision to speed or not to speed is displaced from the moral agency of drivers to that of legislators who decide to implement ISA. In this case, each of the three drivers loses the specific option to exercise one's capacities for self-control in respecting the traffic laws, be it for prudential or moral reasons. Limiting ISA not only rules out the opportunity for drivers to show consideration for the safety of others by not speeding, it also rules out the option for these drivers to engage in deliberation about the moral reasons for speeding in emergency cases. Yeung discusses the case of a driver who sees a collapsing elderly pedestrian who urgently needs medical assistance (Yeung: 2011, 13). Limiting ISA prevents speeding and makes it futile for the driver to make his own judgement as to whether morality requires the transgression of the legal speed limits to rescue the elderly person.

The interpretation of the law by the driver, who is the subject of the law, has become otiose, and the situation in which this driver would have to defend himself for speeding never arises. In the present system, however, if the driver were caught speeding, the legal system would provide several opportunities to account for his behaviour, such as arguing that prosecution was not appropriate or pleading in court that his behaviour was justified or excusable (cf. Yeung: 2011, 13). Nonetheless, the driver in the car with limiting ISA can still deliberate about other options to rescue the wounded pedestrian, such as calling emergency services.

Interestingly, this picture is completely changed by a relatively narrow option to override the limiting ISA system that is restricted in terms of both duration and excess speed. Drivers can exercise their capacities for moral agency by not using the override option; they can decide in emergency situations to speed while being willing to account for their actions, and it is once more reasonable to praise or blame their choice of driving speed.[115] The same applies to a larger extent to supportive ISA (which limits the speed but can be overridden at *any* time).

This more fine-grained analysis of how the different versions of ISA affect our opportunities for moral agency shows that this effect can be justified by the road safety benefits. The upshot of the analysis is that advisory ISA supports drivers in their exercise of moral agency. However, limiting ISA, on the contrary, does reduce drivers' moral agency by making speeding impossible. At the same time, drivers still have plenty of *other* opportunities to harm fellow road users. Drivers can fail to stay in their lane, engage in tailgating, or simply be inattentive. Thus, even limiting ISA does not significantly corrode our moral community and to the small extent that it does, this can be plausibly justified in terms of the avoidance of harm to others.[116] It is informative to compare the use of ISA to that of speed bumps and elevated pavement. The two latter measures also reduce opportunities for moral agency, but seem not to elicit any concerns

[115] The reason to think that overridable limiting ISA system will still be considerably more effective than advisory and supportive ISA is that it forces that group of drivers that ignores advisory ISA and persistently overrides supportive ISA to give up nearly all of their former habit of speeding. An anonymous reviewer suggested a way to severely restrict the override function: allow "the driver [only to] speed after sending a signal to some authority that they are going to speed." A nice feature of an override is also that it will reduce some of the ISA-generated risks discussed in the previous section. See also Yeung (2011), and Brownsword (2005) for some thoughts about an override to techno-regulation.

[116] But perhaps from the subjective point of view of some drivers, it *feels* as if these technologies take away precisely a very important opportunity for moral agency. One they find important and which they enjoy during considerable amounts of time. Limiting ISA denies them the status of responsible and capable agents in a domain they highly value and which in our car culture may be part of their authenticity. I leave it an open question whether the subjective experience of a specific opportunity for moral agency as being important *makes* it a greater contribution to the flourishing of the moral community. However, given the massive occurrence of speeding, many of these drivers just misattribute to themselves this status of responsible driver. Their resistance to limiting ISA is most plausibly interpreted as resistance against lacking the option to speed, and not as resistance against lacking the option to show respect to fellow road users. This interpretation is supported by research that shows that drivers who confess to enjoy speeding also are most likely to override ISA systems (see Jamson: 2006).

from any of the involved parties. In these cases as well, the huge safety benefits have a larger moral weight than considerations of moral agency. To conclude, the objection that ISA is detrimental to driver's moral agency and responsibility is unconvincing.

7.7. Objection III: Accepting ISA leads to a *Brave New World* society

However, there is a related concern that is more pressing: if we accept ISA to gain safety, we seem to commit ourselves to accepting even more techno-regulation for similar reasons, ultimately leading to erosion of moral agency and responsibility. Brownsword rightly points to the fact that the effectiveness of techno-regulation is a strong incentive for regulators to apply such regulation in many domains (Brownsword: 2005, 19). Citizens would then perceive that their moral responsibility was being displaced to the system, and their capacities for moral agency would become superfluous in many situations and weakened as a result. Ultimately, this situation would lead to a society such as that in *Brave New World* in which the government extensively uses technology to 'design out' all socially undesirable behaviour.

It is not inevitable or even likely that accepting (limiting) ISA would lead to such a dreadful society. Yeung develops key insights into why this is unlikely (2011). Most importantly, society should maintain *sufficient* opportunities for the right type of moral agency to sustain the moral community. Yeung argues that we should only accept techno-regulation to an extent that is compatible with maintaining a healthy moral community. Crucially, new technologies always introduce new options for moral agency and therefore always affect the health of the moral community. This effect depends on the extent and types of these opportunities that come into existence with new technologies and disappear with displaced old technologies.

If we agree that what matters is having sufficient opportunities for moral agency, we can see that maximising these opportunities will often come at too high a cost. For example, faster cars increase our opportunity to exercise moral restraint and respect fellow humans. However, the other side of the coin is an increased opportunity to harm these humans. In current practice, faster cars increase the number of people killed, thereby removing members of the very same moral community. Several trade-offs can be identified in this respect. Maximising *opportunities* for moral agency does not maximise the development of our *capacities* for moral agency. Developing these capacities is difficult in a

Hobbesian state of nature because acting on moral reasons, in principle, is most likely to occur in circumstances under which mutual trust exists among members of a society. This trust can only be established in a state that secures a minimum amount of safety to life and limb to which techno-regulation may be well suited. A second trade-off occurs between the opportunities for moral agency of different agents. If, for example, cars become faster and consequently drivers can do more harm with their cars, others may feel that it is no longer a responsible choice to walk or cycle.[117]

Once we accept this perspective on techno-regulation, we need to find a principled way to determine which instances we will allow and not allow. Our aim should be to maintain a sufficient opportunity for moral agency to sustain the moral community. Because techno-regulation is a relatively new phenomenon, legal scholars have just started to develop frameworks for assessing techno-regulation that could serve this purpose. Yeung provides some considerations that a fully worked out framework would be likely to incorporate (2011, 23–27). I will apply these considerations to ISA technology to estimate whether ISA will be one of the pieces of techno-regulation that societies are likely to accept. First, the regulatory purpose must be legitimate, and the social benefits of the measure must be substantial. Regarding ISA, the regulatory purpose of reducing injury and lethal accidents is clearly legitimate, and the expected social benefits are enormous, both in terms of lives saved, which are ultimately valuable, and prevented economic and social losses.

Second, the effect on the moral community of reducing the number of options for moral agency must be justified by these social benefits. In the previous section, I showed that ISA has a minor effect in this respect that is clearly outweighed by the social benefits. To see this more clearly, compare ISA with, for example, devices that ensure that only paying passengers can use public transport. These devices take away a significant opportunity to act honestly, and the long-term social costs of reinforcing a ‘pay only if you are forced to pay’ attitude may be high. However, the benefits of such devices are only financial and do not involve preventing the loss of life and limb.

Third, the technological measures must not be harmful to the regulatees or be otherwise illegitimate. This partly depends on the ethical and democratic standards of the society. ISA also satisfies this third criterion. ISA differs from

[117] I thank Auke Pols for this point.

many other measures that would do equally well qua trading moral agency options for large social benefits. ISA targets what people can do with a technological artefact, whereas other measures target persons themselves and may compromise their rights. For example, tagging or chipping former criminals who have completed their sentences to prevent recidivism would, given substantial recidivism, result in significant social benefits (cf. Brownsword & Goodwin: 2012). However, such measures go against the maxim that former criminals are citizens whose rights have been fully reinstated and can seriously harm these former criminals. Being trusted and having privacy is a crucial condition for a former criminal to resume his life as a citizen. ISA does not violate any drivers' rights or otherwise treat them illegitimately.

Although even an established and well-functioning framework for deciding on techno-regulation may give rise to considerable debate and judgements concerning degrees, Yeung's three considerations give reason to think that limiting ISA will lie on the justified side of the spectrum. I expect that satisfactory frameworks for techno-regulation can be developed, just like our society has developed principled ways to decide which behaviours to criminalise. Accepting ISA will not commit us to a world replete with techno-regulation.

7.8. Concluding remarks

I have argued that obligatory ISA is justified on the basis of its considerable potential to prevent harm to other road users caused by speeding. Exceeding the speed limits imposes significant excess risks as compared to lawful driving. These risks cannot be justified, because the interests of other road users in the safety of life and limb outweigh the interests of drivers in further time saving and driving pleasure. My argument that even lawful driving involves morally problematic risks confers additional justification to ISA. It is worth mentioning here that ISA also leads to saving a few percent of fuel, depending on circumstances (Lai et al.: 2012). Three specific objections against ISA have been extensively evaluated and shown to be ultimately unconvincing.

Although all versions of ISA are justified by the harm they prevent and survive the objections, it is still the question for which the strongest moral case can be made. The answer depends on weighing the various considerations discussed above. First, the effectiveness of advisory ISA, although sufficient to lead to a positive cost-benefit analysis (Lai et al.: 2012), is low compared to supportive ISA and particularly so to limiting ISA. Limiting ISA is approximately three to ten times more effective in reducing accidents than advisory ISA (section 7.3), depending on which studies are used to perform the comparison. Safety benefits are crucial to all of the aforementioned arguments; thus, the higher effectiveness of limiting ISA over that of advisory ISA is a major advantage. However, advisory ISA performs better for all other criteria: advisory ISA does not reduce but supports drivers' moral agency, introduces fewer additional safety risks to drivers, and, by virtue of its larger acceptance by society, adheres more strongly to the ideal that lawmaking and law enforcement are a cooperative enterprise of citizens and government (legality). Nonetheless, if we view the effort to increase road safety as a task for all citizens, it is unfair that drivers who comply with advisory (and supportive) ISA must bear the risks imposed by non-compliant drivers. That is, limiting ISA reduces the unfairness of the present mutual imposition of risks between lawful and non-lawful drivers more significantly than advisory ISA.

I take the impressive effectiveness of limiting ISA as being decisive for considering its moral case to be the strongest. Nonetheless, sufficient public support for ISA is essential. A strictly limited override will facilitate this objective and has also been shown to improve the performance of limiting ISA on driver safety and the preservation of moral agency.

However, making a final judgement of this type at this point in time is only of academic value. Governments that decide to implement ISA should consider an implementation trajectory, such as that proposed by Carsten and Tate (2005). This trajectory starts with self-chosen advisory ISA and proceeds via several steps to eventual obligatory limiting ISA in all cars. These steps may include starting with subclasses of drivers, such as young adults or repeat offenders. In the course of such a trajectory, societal acceptance and democratic support will most likely grow, while manufacturers will gain more knowledge about the risks associated with ISA, which will help improve the technology and lead to a more informed implementation process. The technology needed for advisory ISA is already widely available in the form of navigation devices and smart phones (O. Carsten: 2012). Governments have ample moral reasons to start implementing advisory ISA today.[118]

[118] I am grateful to Jan-Willem van der Pas, Oliver Carsten, and Niels Bos for help with the technical details of ISA.

8 The ethics of accident-algorithms for self-driving cars: An applied trolley problem?[119]

8.1. Introduction

Self-driving cars hold out the promise of being much safer than our current manually driven cars. This is one of the reasons why many are excited about the development and introduction of self-driving cars. Yet, self-driving cars cannot be a 100% safe. This is because they will drive with high speed in the midst of unpredictable pedestrians, bicyclists, and human drivers (Goodal: 2014a-b). So there is a need to think about how they should be programmed to react in different scenarios in which accidents are highly likely or unavoidable. This raises important ethical questions. For instance, should autonomous vehicles be programmed to always minimize the number of deaths? Or should they perhaps be programmed to save their passengers at all costs? What moral principles should serve as the basis for these "accident-algorithms"? Philosophers are slowly but surely beginning to think about this general issue, and it is already being discussed in the media and in various different online forums.

Some philosophers have recently likened accident-management in autonomous vehicles to the so-called trolley problem. Several journalists and opinion piece writers have also done so.[120] The trolley problem is the much-discussed set of philosophical thought experiments in which there is a runaway trolley and the only way to save five people on the tracks is to sacrifice one person (Thomson: 1985b). Different versions of these trolley cases vary with respect to how the one will need to be sacrificed in order for the five to be saved. It is the most basic versions that are said to foreshadow the topic of how to program autonomous vehicles.

[119] This chapter has been published in *Ethical Theory and Moral Practice* (Nyholm & Smids: 2016)

[120] E.g., (Achenbach: 2015; Doctorow: 2015; Lin: 2013; Windsor: 2015; Worstall: 2014).

For example, Patrick Lin writes:

> One of the most iconic thought-experiments in ethics is the trolley problem . . . and this is one that may now occur in the real world, if autonomous vehicles come to be (Lin: 2015).

Similarly, when discussing another kind of autonomous vehicles (viz. driverless trains), Wendell Wallach and Colin Allen write:

> . . . could trolley cases be one of the first frontiers for artificial morality? Driverless systems put machines in the position of making split-second decisions that could have life or death implications. As the complexity [of the traffic] increases, the likelihood of dilemmas that are similar to the basic trolley case also goes up (Wallach & Allen: 2009, 14).

Nor are philosophers alone in making this comparison. Economists and psychologists Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan write:

> situations of unavoidable harms, as illustrated in [our examples of crashes with self-driving cars], bear a striking resemblance with the flagship dilemmas of experimental ethics – that is, the so-called 'trolley problem' (Bonnefon, Shariff, & Rahwan: 2015, 3).

According to these various writers, then, the problem of how to program self-driving cars and other autonomous vehicles for different accident-scenarios is very similar to the trolley problem. If true, this would have important implications concerning how to best approach the ethics of self-driving cars. It suggests that when we approach the ethics of accident-algorithms for autonomous vehicles, the ever growing literature on the trolley problem is a good, if not the best, place to start. Moreover, it suggests that that literature treats the key issues we need to focus on when we try to formulate an ethical framework for sound moral reasoning about how autonomous vehicles should be programmed to deal with risky situations and unavoidable accidents.[121]

[121] Reasoning in just that way, Bonnefon et al. (2015) propose that the methods developed within experimental ethics to investigate judgments about the trolley problem should be used to investigate ordinary people's intuitions about accident-algorithms for self-driving cars. Bonnefon et al. further think that once these intuitions have been carefully surveyed and systematically analyzed, they should then serve as starting points for normative discussions of how self-driving cars ought to be programmed.

In this chapter, we critically examine this tempting analogy between the trolley problem and the problem of accident-algorithms for self-driving cars. We do so with a skeptical eye. Specifically, we argue that there are three very important respects in which these two topics are not analogous. We think, therefore, that it is important to resist the temptation to draw a very strong analogy between the ethics of accident-algorithms for self-driving cars and the philosophy of the trolley problem.

Why is this an important topic to investigate? Firstly, the issue of how to program self-driving cars is a pressing ethical issue, given the rapid development of this technology and the serious risks involved. We therefore need to identify the best sources of ethical theory that could help us to deal with this part of moral practice. At this stage, we are only beginning to grapple with this problem. So it is crucial to thoroughly investigate any initial "leads" we have about where best to start as we open up the discussion of this general topic. The similarity between accident-planning for self-driving cars and the trolley problem that some writers claim to have identified is one such lead. That's one reason why it is important to investigate whether (or not) the literature on the trolley problem is indeed the best place to primarily turn to as we approach this ethical issue. Secondly, in investigating how similar or dissimilar these two topics are, we in effect isolate and identify a number of basic key issues that further work on the ethics of accident-algorithms for self-driving cars needs to deal with. In conducting this positive part of our inquiry, we investigate what types or categories of considerations are at issue here. And we make it very clear that the problem of how to program autonomous vehicles to respond to accident-scenarios is a highly complex ethical issue, under which there are various sub-issues that on their own also exhibit a lot of complexity.

We proceed as follows. We first say a little more about why an ethical framework for risk-management for autonomous vehicles needs to be developed (section 8.2). We then say more about the trolley problem and the main issues discussed in the literature on the trolley problem (section 8.3). After that, we explain the three main differences we see between the ethics of accident-management in self-driving cars, on the one hand, and the trolley problem on the other hand (sections 8.4-6). To anticipate, these differences have to do with (i) prospective planning by groups that takes large numbers of situational features into account vs. imagined split-second decisions by individuals that abstract away all but a few features of the situation directly at hand; (ii) taking seriously pressing issues of moral and legal responsibility vs. setting such issues aside as irrelevant as a matter of stipulation; and (iii) reasoning about probabilities, uncertainties and risk-management vs. intuitive judgments about what are stipulated to be known and fully certain facts. Lastly, we end by briefly summarizing our main conclusions (section 8.7).

8.2. Programming Self-Driving Cars for How to React in the Event of Accidents

As we noted above, even self-driving cars will inevitably sometimes crash. Noah Goodall (2014a-b) and Patrick Lin (2015) convincingly argue for this claim, and in addition explain why programming self-driving cars for crashes involves ethical choices. In explaining what's at issue, we here draw on Goodall's and Lin's work.

A self-driving car uses advanced sensor-technology to detect its surroundings and sophisticated algorithms to subsequently predict the trajectory of nearby (moving) objects. Self-driving cars can also use information technology to communicate with each other, thus achieving better coordination among different vehicles on the road. However, since cars are heavy and move with high speed, physics informs us that they have limited maneuverability, and that they often cannot simply stop. Therefore, even if the car-to-car communication and the sensors and algorithms are all functioning properly (and would be better than current technology), self-driving cars will not always have sufficient time to avoid collisions with objects that suddenly change direction (Goodall: 2014a). Self-driving cars will sometimes collide with each other. But there are also other moving objects to worry about. Pedestrians, cyclists, and wildlife naturally come to mind here (Lin: 2015). However, we must also take into account human-driven cars. Because, as is generally acknowledged by the experts, self-driving cars will for a long period drive alongside human-driven cars (so-called "mixed traffic", see (van Loon & Martens: 2015)).

For these reasons, automated vehicles need to be programmed for how to respond to situations where a collision is unavoidable; they need, as we might put it, to be programmed for how to crash. At first blush, it might seem like a good idea to always transfer control to the people in the car in any and all situations where accidents are likely or unavoidable. However, human reaction-times are slow. It takes a relatively long time for us to switch from focusing on one thing to focusing on another. So handing over control to the human passengers will often not be a good option for the autonomous vehicle. Hence the car itself needs to be prepared, viz. programmed, for how to handle crashes.

This has certain advantages. A self-driving car will not react in the panicky and disorganized ways a human being is apt to react to accident-scenarios. Even in situations of unavoidable collisions, the car's technology enables significant choices and control-levels regarding how to crash. Based on its sensor inputs and the other information it has access to, the car can calculate the most likely consequences of different trajectories that involve different combinations of braking and swerving.

Consider now the following scenario.[122] A self-driving car with five passengers approaches a conventional car (e.g. a heavy truck) that for some reason suddenly departs from its lane and heads directly towards the self-driving car. In a split-second, the self-driving car senses the trajectory and the likely weight of the oncoming truck. It calculates that a high-impact collision is inevitable, which would kill the five passengers, unless the car swerves towards the pavement on its right-hand side. There, unfortunately, an elderly pedestrian happens to be walking, and he will die as a result if the self-driving car swerves to the right and hits him. This is the sort of situation in which the human passengers of a self-driving car cannot take control quickly enough. So the car itself needs to respond to the situation at hand. And in order for the five passengers in the self-driving car to be saved, as they are likely to be if the head-on collision with the heavy truck is avoided, the car here needs to make a manoeuvre that will most likely kill one person.

It is evident that scenarios like this one involve significant ethical dilemmas. Among other things, they raise questions about what the self-driving car's pre-set priorities should be. Should it here swerve to the sidewalk and save the greatest number, or should it rather protect the innocent pedestrian and crash into the oncoming truck? In general, should the car be programmed to always prioritize the safety of its passengers, or should it sometimes instead prioritize other considerations, such as fairness or the overall good, impartially con-

[122] Our illustration here is a "mixed traffic"-case. Self-driving cars will inevitably sometimes collide with each other, for example if one of them is malfunctioning. But the risks are even greater within mixed traffic involving both self-driving cars and conventional cars, since human drivers and self-driving cars have a harder time communicating with each other (van Loon and Martens: 2015). Still, self-driving cars need to be programmed for how to handle collisions with both other self-driving cars and conventional cars (in addition to any other objects that might suddenly appear in their paths (Lin: 2015)). We discuss the ethics of compatibility-problems within mixed traffic at greater length in (Nyholm & Smids: forthcoming.).

sidered? Crucially, unless the self-driving car is programmed to respond in determinate ways to morally loaded situations like the one we just described, there is an unacceptable omission in its readiness to deal with the realities and contingencies of actual traffic (Goodall: 2014a-b; Lin: 2015). Not programming the car for how to respond to situations like this and others like it amounts to knowingly relinquishing the important responsibility we have to try to control what happens in traffic. It amounts to unjustifiably ignoring the moral duty to try to make sure that things happen in good and justifiable ways. We should not do that. Hence the need for ethical accident-algorithms.[123]

At first glance, the accident-scenario we just described above looks similar to the examples most commonly discussed in relation to the trolley problem. But suppose that we probe a little deeper beneath the immediate surface, and that we home in on the more substantial ethical issues raised by the choice of accident-algorithms for self-driving cars. Are Lin, Wallach and Allen, and Bonnefon et al. then still right that the choice of accident-algorithms for self-driving cars is like a real world version of the trolley problem, as it is usually understood and discussed in the literature?

8.3. “The Trolley Problem”?

The easiest way to introduce the trolley problem is to start with the two most widely discussed cases involved in these thought experiments. These are the cases the above-cited writers have in mind when they compare the ethics of accident-algorithms for self-driving cars to the trolley problem.

In the “switch” case, a driverless trolley is heading towards five people who are stuck on the tracks and who will be killed unless the trolley is redirected to a side track. You are standing next to a switch. If you pull the switch, the trolley is redirected to a side-track and diverted away from the five. The trouble is that on this side track, there is another person, and this person will be killed if you pull

[123] The design of ethical decision-making software immediately presents two major challenges. First, what moral principles should be employed to solve this sort of ethical dilemmas? Second, even if we were to reach agreement, it turns out to be a formidable challenge to design a car capable of acting fully autonomously on the basis of these moral principles (cf. Goodall: 2014a). We will not go into this latter question, but will instead here simply note that this is an important and pressing issue, which is studied in the field of machine morality or robot ethics (e.g. Wallach and Allen: 2009).

the switch to redirect the train. Nevertheless, a very common response to this case is that it is here permissible for you to save the five by redirecting the train, thus killing the one as a result (Greene: 2013).

In a different variation on this theme, the "footbridge" case, saving the five requires different means. In this case, you are located on a footbridge over the tracks. Also present on the footbridge is a very large and heavy man. His body mass is substantial enough that it would stop the trolley if he were pushed off the footbridge and onto the tracks. But this would kill him. Is it morally permissible to push this man to his death, thereby saving the five by this means? A very common response to this case is that it is not permissible (Greene: 2013). So in this case saving the five by sacrificing the one seems wrong to most of us, whereas in the other case, saving the five by sacrificing the one seems morally permissible.

Many people casually use the phrase "the trolley problem" to refer to one or both of these examples. But some influential philosophers use this phrase to mean something more distinct. According to Judith Jarvis Thomson, for example, the basic trolley problem is to explain the above-described asymmetry in our judgments (Thomson: 2008). That is, why is it permissible in one case to save the five by sacrificing the one, whereas it is not permissible to save the five by sacrificing the one in the other case? Others favor a wider interpretation of the trolley problem, holding that this problem also arises in cases that don't involve any trolleys at all. According to Frances Kamm, the basic philosophical problem is this: why are certain people, using certain methods, morally permitted to kill a smaller number of people to save a greater number, whereas others, using other methods, are not morally permitted to kill the same smaller number to save the same greater number of people (Kamm: 2015)? For example, why is it is not permissible for a medical doctor to save five patients in need to organ-transplants by "harvesting" five organs from a perfectly healthy patient who just came into the hospital for a routine check-up? This case doesn't mention trolleys, but Kamm thinks it nevertheless falls under the wide umbrella of the trolley problem.

These various thought-experiments have been used to investigate a number of different normative issues. For example, they have been used to investigate the difference between: (i) "positive" and "negative" duties, that is, duties to do certain things vs. duties to abstain from certain things; (ii) killing and letting die; and (iii) consequentialism and non-consequentialism in moral theory, that is, the difference between moral theories only concerned with promoting the overall good vs. moral theories that also take other kinds of considerations into account (Foot: 1967; Kamm: 2015; Thomson: 1985b). And in recent years, they have also been used to empirically investigate the psychology and neuroscience of different types of moral judgments (Greene: 2013; Mikhail: 2013).

We agree that trolley cases can certainly be useful in the discussion of these various topics. But how helpful are these thought-experiments and the large literature based on them for the topic of how self-driving cars ought to be programmed to respond to accident-scenarios where dangerous collisions are highly likely or unavoidable? We will now argue that there are three crucial areas of disanalogy that should lead us to resist the temptation to draw a strong analogy between the trolley problem and the ethics of accident-algorithms for self-driving cars.

8.4. Two Very Different Decision-Making Situations

To explain the first noteworthy difference between the ethics of accident-algorithms for autonomous vehicles and the trolley dilemmas we wish to bring attention to, we will start by returning to the quote from Wallach and Allen in the introduction above. Specifically, we would like to zoom in on the following part of that quote:

> Driverless systems put machines in the position of making split-second decisions that could have life or death implications.

It is tempting to put things like this if one wishes to quickly explain the ethical issues involved in having driverless systems in traffic.[124] But it is also somewhat misleading. Wallach and Allen are surely right that there is some sense in which the driverless systems themselves need to make "split-second decisions" when they are in traffic. And these can indeed have life or death implications. However, strictly speaking, the morally most important decision-making is made at an earlier stage. It is made at the planning stage when it is decided how the autonomous vehicles are going to be programmed to respond to accident-

[124] In a similar way, Lin writes that if "motor vehicles are to be truly autonomous and be able to operate responsibly on our roads, they will need to replicate [. . .] the human decision-making process." (Lin: 2015, 69). Cf. also Purves et al.'s remark that "[d]riverless cars [. . .] would likely be required to make life and death decisions in the course of operation" (Purves, Jenkins, & Strawser: 2015, 855).

scenarios. The “decisions” made by the self-driving cars implement these earlier decisions.

The morally relevant decisions are prospective decisions, or contingency-planning, on the part of human beings. In contrast, in the trolley cases, a person is imagined to be in the situation as it is happening. The person is forced right there and then to decide on the spot what to do: to turn the trolley by pulling the switch (switch case) or push the large man off the bridge (footbridge case). This is split-second decision-making. It is unlike the prospective decision-making, or contingency-planning, we need to engage in when we think about how autonomous cars should be programmed to respond to different types of scenarios we think may arise.[125] When it comes to the morally relevant decision-making situations, there is more similarity between accident-situations involving conventional cars and trolley-situations than between prospective programming for accident-situations involving autonomous cars and trolley-situations. For example, a driver of a conventional car might suddenly face a situation where she needs to decide, right there and then, whether to swerve into one person in order to avoid driving into five people. That is much closer to a trolley-situation than the situation faced by those who are creating contingency-plans and accident-algorithms for self-driving cars is.[126]

Nor is it plausible to think of decision-making about how self-driving cars should be programmed as being made by any single human being. That is what we imagine when we consider the predicament of somebody facing a trolley situation. This does not carry over to the case of self-driving cars. Rather, the decision-making about self-driving cars is more realistically represented as being made by multiple stakeholders – for example, ordinary citizens, lawyers, ethicists, engineers, risk-assessment experts, car-manufacturers, etc. These stakeholders need to negotiate a mutually agreed-upon solution. And the agreed-

[125] As an anonymous reviewer pointed out, this difference in time-perspectives might render it plausible for different moral principles to serve as the evaluation-criteria for the programming of self-driving cars and the behavior of drivers in acute situations. For example, the aim of minimizing the statistically expected number of deaths can seem more justifiable and apt in prospective decision-making about accident-algorithms for self-driving cars than in retrospective evaluation of actual human responses to dramatic accident-scenarios (Hansson: 2013, 74–80).

[126] We owe this last observation to our colleague Auke Pols.

upon solution needs to be reached in light of various different interests and values that the different stakeholders want to bring to bear on the decision.[127]

The situation faced by the person in the trolley case almost has the character of being made behind a "veil of ignorance," in John Rawls' terms (Rawls: 1971). There is only a very limited number of considerations that are allowed to be taken into account. The decision-maker is permitted to know that there are five people on the tracks, and that the only way to save them is to sacrifice one other person – either by redirecting the runaway trolley towards the one (switch case) or by pushing a large person into the path of the trolley (footbridge case). These are the only situational factors that are allowed into the decision-making, as if this were a trial where the jury is only allowed to take into account an extremely limited amount of evidence in their deliberations.

This is not the ethical decision-making situation that is faced by the multiple stakeholders who together need to decide how to program self-driving cars to respond to different types of accident-scenarios. They are not in a position where it makes sense to set aside most situational and contextual factors, and only focus on a small set of features of the immediate situation. Instead, they can bring any and all considerations they are able to think of as being morally relevant to bear on their decisions about how to program the cars. They can do that and should do so.

In sum, the basic features of these two different decision-making situations are radically different. In one case, the morally relevant decision-making is made by multiple stakeholders, who are making a prospective decision about how a certain kind of technology should be programmed to respond to situations it might encounter. And there are no limits on what considerations, or what numbers of considerations, might be brought to bear on this decision. In the other case, the morally relevant decision-making is done by a single agent who is responding to the immediate situation he or she is facing – and only a very limited number of considerations are taken into account. That these two deci-

[127] Jason Millar argues that the accident-algorithms of self-driving cars ought to be selected by the owner of the car (Millar: 2014). This would mean that different cars could have different accident-algorithms. Two comments: firstly, this would still require a mutual decision since the basic decision to give owners of self-driving cars the right to choose their accident-algorithms would need to be agreed upon by the various different stakeholders involved. Second, this seems undesirable since different accident-algorithms in different cars would complicate coordination and compromise safety.

sion-making situations are so radically different in their basic features in these respects is the first major disanalogy we wish to highlight.[128]

8.5. A Second Disanalogy: The Importance of Moral and Legal Responsibility

In order to set up the second main observation we wish to make, we will start by again returning to the following feature of the standard trolley cases. As just noted, we are asked to abstract away all other aspects of the situations at hand except for the stipulation that either five will die or one will die, where this depends on whether we (i) redirect the train away from the five and towards the one by pulling a switch (switch case) or (ii) push the one down from the bridge onto the tracks and into the line of the trolley (footbridge case). We are supposed to bracket any and all other considerations that might possibly be ethically relevant and consider what it would be permissible or right to do, taking only these features of the cases into consideration.

This is a characteristic of the trolley cases that has been criticized. Consider Allen Wood's criticism. Wood notes that, when we set aside everything else than the above-described considerations, there is "an important range of considerations that are, or should be, and in real life would be absolutely decisive in our moral thinking about these cases in the real world is systematically abstracted out" (Wood: 2011, 70). Explaining what he finds problematic about this, Wood writes:

> even if some choices [in real life] inevitably have the consequence that either one will die or five will die, there is nearly always something wrong with looking at the choice *only* in that way. (Wood: 2011, 73, emphasis added)

What is Wood getting at? What Wood is missing is, among other things, due concern with moral and legal responsibility, viz. the question of who we can

[128] Jan Gogoll and Julian Mueller identify three further differences between these two decision-making situations worth noting: (i) the much more static nature of standard trolley-situations as compared to the non-static situations that self-driving cars will typically face; (ii) the possibility of updating and revising accident-algorithms over time in self-driving cars, which contrasts with how trolley-situations are typically represented as isolated events; (iii) the one-sided nature of the "threat" in trolley-situations (the decision-maker is not represented as being at risk) as opposed to how in typical traffic-situations, all parties are usually subject to certain risks (Gogoll & Müller: 2016).

justifiably hold morally and legally responsible for what is going on. Commenting specifically on how trains and trolley cars are regulated in real life, Wood writes:

> Trains and trolley cars are either the responsibility of public agencies or private companies that ought to be, and usually are, carefully regulated by the state with a view to ensuring public safety and avoiding loss of life (Wood: 2011, 74).

Developing the legal side of the issue further, Wood continues:

> . . . mere bystanders ought to be, and usually are, physically prevented from getting at the switching points of a train or trolleys. They would be strictly forbidden by law from meddling with such equipment for any reason, and be held criminally responsible for any death or injury they cause through such meddling (Wood: 2011, 75).

Thus Wood thinks that trolley cases are too far removed from real life to be useful for moral philosophy. One key reason is that in real life, we hold each other responsible for what we do or fail to do. When it comes to things involving substantial risks – such as traffic – we cannot discuss the ethical issues involved without taking issues of moral and legal responsibilities into account. Since trolley cases ignore all such matters, Wood finds them irrelevant to the ethics of the real world.

We think that Wood might be going too far in making this criticism of the philosophical and psychological literature on the trolley problem. It is surely the case that sometimes examples that might not be very true to real life can serve useful purposes in moral philosophy and in various other fields of academic inquiry. But we think that the issue Wood brings up helps to highlight a stark difference between the discussion of the trolley problem and the issue of how we ought to program self-driving cars and other autonomous vehicles to respond to high-risk situations.

The point here is that when it comes to the real world issue of the introduction of self-driving cars into real world traffic, we cannot do what those who discuss the trolley problem do. We cannot stipulate away all considerations having to do with moral and legal responsibility. We must instead treat the question of how self-driving cars ought to be pre-programmed as partly being a matter of what people can be held morally and legally responsible for (cf. Hevelke & Nida-Rümelin: 2014). Specifically, we must treat it as a question of what those who sell or use self-driving cars can be held responsible for and what the society that permits them on its roads must assume responsibility for.

With the occurrence of serious crashes and collisions – especially if they involve fatalities or serious injuries – people are disposed to want to find some person or persons who can be held responsible, both morally and legally. This is an unavoidable aspect of human interaction. It applies both to standard traffic, and traffic introducing self-driving cars. Suppose, for example, there is a collision between an autonomous car and a conventional car, and though nobody dies, people in both cars are seriously injured. This will surely not only be followed by legal proceedings. It will also naturally – and sensibly – lead to a debate about who is morally responsible for what occurred. If the parties involved are good and reasonable people, they themselves will wonder if what happened was "their fault". And so we need to reflect carefully on, and try to reach agreement about, *what* people can and cannot be held morally and legally responsible for when it comes to accidents involving self-driving cars. We also need to reflect on, and try to reach agreement about, *who* can be held responsible for the things that might happen and the harms and deaths that might occur in traffic involving these kinds of vehicles.

Questions concerning both "forward-looking" and "backward-looking" responsibility arise here. Forward-looking responsibility is the responsibility that people can have to try to shape what happens in the near or distant future in certain ways. Backward-looking responsibility is the responsibility that people can have for what has happened in the past, either because of what they have done or what they have allowed to happen (van de Poel: 2011). Applied to risk-management and the choice of accident-algorithms for self-driving cars, both kinds of responsibility are highly relevant. One set of important questions here concerns moral and legal responsibility for how cars that will be introduced into traffic are to be programmed to deal with the various different kinds of risky situations they might encounter in traffic. Another set of questions concerns who exactly should be held responsible, and for what exactly they should be held responsible, if and when accidents occur. The former set of questions are about forward-looking responsibility, the second about backward-looking responsibility. Both sets of questions are crucial elements of the ethics of self-driving cars.[129]

We will not delve into how to answer these difficult questions about moral and legal responsibility here. Our point in the present context is rather that these are pressing questions we cannot ignore, but must instead necessarily grapple with when it comes to the ethics of accident-algorithms for self-driving cars. Such questions concerning moral and legal responsibility are typically simply set aside in discussions of the trolley problem. For some of the theoretical purposes the trolley cases are meant to serve, it might be perfectly justifiable to do so. In contrast, it is not justifiable to set aside basic questions of moral and legal responsibility when we are dealing with accident-algorithms for self-driving cars. So we here have a second very important disanalogy between these two topics.

8.6. Stipulated facts and certainties vs. risks, probabilities, and uncertainties

We now turn to the third and last major disanalogy we wish to highlight. Here, too, we will approach this disanalogy via a criticism that has been raised against the trolley problem and its relevance to the ethical issues we face in the real world. What we have in mind is Sven Ove Hansson's criticism of standard moral theory and what he regards as its inability to properly deal with the risks and uncertainties involved in many real world ethical issues. In one of his recent papers on this general topic, Hansson specifically brings up the trolley problem as one clear case in point of what he has in mind. Hansson writes:

> The exclusion of risk-taking from consideration in most of moral theory can be clearly seen from the deterministic assumptions commonly made in the standard type of life-or-death examples that are used to explore the implications of moral theories. In the famous trolley problem, you are assumed to know that if you flip the switch, then one person will be killed, whereas if you don't flip it, then five other persons will be killed (Hansson: 2012, 44).

[129] Current practice typically assigns backward-looking responsibility for accidents to drivers. But the introduction of self-driving cars is likely to shift backward-looking responsibility-attributions towards car-manufacturers. If justified, this would make backward- and forward-looking responsibility for accidents more closely related and coordinated. We owe these observations to an anonymous reviewer.

What is Hansson's worry here? What is wrong with being asked to stipulate that we know the facts and that there is no uncertainty as regards what will happen in the different sequences of events we could initiate? Hansson comments on this aspect of the trolley cases in the following way:

> This is in stark contrast to ethical quandaries in real life, where action problems with human lives at stake seldom come with certain knowledge of the consequences of the alternative courses of action. Instead, uncertainty about the consequences of one's actions is a major complicating factor in most real-life dilemmas (op cit.).

Hansson is not alone in making this criticism. Others have also worried that it is absurd to suppose, in any realistic situation, that doing something such as to push a large person in front of a trolley car would be sure to stop the trolley and save any people who might be on the tracks. As before, however, this may not be a fatal objection to the trolley cases if we conceive of them as a set of stylized thought experiments we use for certain circumscribed purely theoretical and abstract purposes. But again, we also see here that the trolley cases are far removed from the reality that we face when we turn to the ethical problem of how to program self-driving cars to respond to risky situations when we introduce these cars into actual traffic and thereby bring them into the real world with all its messiness and uncertainty.

We will illustrate this point by taking a closer look at our scenario from section I above. This was the scenario in which a heavy truck suddenly appears in the path of a self-driving car carrying five passengers, and in which the only way for the self-driving car to save the five appeared to be to swerve to the right, where it would kill an elderly pedestrian on the sidewalk. Under this brief description, the scenario might appear to involve an ethical dilemma in which we need to choose between outcomes whose features are known with certainty. But once we add more details, it becomes clear that there is bound to be a lot of uncertainty involved in a more fully described and maximally realistic version of the case.

First, the self-driving car cannot acquire certain knowledge about the truck's trajectory, its speed at the time of collision, and its actual weight. This creates uncertainty because each of these factors has a strong causal influence on the fatality risk for the passengers of the self-driving car (Berends: 2009; Evans: 2001). Moreover, the truck-driver might try to prevent the accident by steering back to her lane. Or if it's already too late, she might start braking just half a second before the crash (thereby significantly reducing the truck's speed and impact). The self-driving car's software can only work with estimates of these alternative courses of events.[130]

Second, focusing on the self-driving car itself, in order to calculate the optimal trajectory, the self-driving car needs (among other things) to have perfect knowledge of the state of the road, since any slipperiness of the road limits its maximal deceleration. But even very good data from advanced sensors can only yield estimates of the road's exact condition. Moreover, regarding each of the five passengers: their chances of surviving the head-on collision with the truck depends on many factors, for example their age, whether they are wearing seat belts, whether they are drunk or not, and their overall state of health (Evans: 2008). The car's technology might enable it to gather partial, but by no means full, information about these issues.[131]

Finally, if we turn to the elderly pedestrian, again we can easily identify a number of sources of uncertainty. Using facial recognition software, the self-driving car can perhaps estimate his age with some degree of precision and confidence (Goodall: 2014a). But it may merely guess his actual state of health and overall physical robustness.[132] And whereas statistical fatality rates for car-pedestrian collisions apply to a whole population, these might ultimately have fairly low predictive value for the elderly pedestrian's more precise chances of survival.[133] Of course, in real life, the scenario also involves the possibility that the pedestrian might avoid being hit by quickly stepping out of the self-driving car's path. The self-driving car necessarily has to work with an estimate of what the pedestrian is likely to do. And this estimation may need to be based on simulation-experiments rather than actual statistics.

As we start filling in these various further details, it quickly becomes clear that what we are dealing with here are not outcomes whose features are known with certainty. We are rather dealing with plenty of uncertainty and numerous

[130] Furthermore, while the self-driving car may recognize the truck-type and know its empty mass, the truck may carry a load whose weight is unknown to the self-driving car.

[131] There is, of course, also the question of whether these kinds of facts about the passengers should count ethically here and if so, how exactly (cf. Lin: 2015)?

[132] It should be noted here that it is controversial whether we should assign any ethical weight to the fact that an elderly person might have a lower chance of surviving an accident than a younger, less fragile person might have. We are not taking a stand on that issue here.

[133] Research on pedestrian fatality rates is still in progress (Rosén, Stigson, & Sander: 2011).

more or less confident risk-assessments (cf. Goodall 2014b, 96). This means that we need to approach the ethics of self-driving cars using a type of moral reasoning we don't have occasion or reason to use in thinking about the standard cases discussed in the trolley problem literature.

In the former case, we need to engage in moral reasoning about risks and risk-management. We also need to engage in moral reasoning about decisions under uncertainty. In contrast, the moral reasoning that somebody facing a trolley case uses is not about risks and how to respond to different risks. Nor is it about how to make decisions in the face of uncertainty. This is a categorical difference between trolley-ethics and the ethics of accident-algorithms for self-driving cars. Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes. The key concepts used differ drastically in what inferences they warrant. And what we pick out using these concepts are things within different metaphysical categories, with differing modal status (e.g. risks of harm, on one side, versus actual harms, on the other).[134]

Thus the distinctive and difficult ethical questions that risks and uncertainty give rise to are not in play in the trolley cases. But they certainly are in the ethics of self-driving cars. Let us give just one illustration. A significant number of people may find hitting the pedestrian morally unacceptable if this was certain to kill him (cf. Thomson: 2008). But what if the estimated chance of a fatal collision were 10 %? Or just 1 %? To many people, imposing a 1% chance of death on an innocent pedestrian in order to save five car-passengers might appear to be the morally right choice. The trolley cases don't require any such judgments. In the scenarios involved in the trolley cases, all outcomes are assumed to be a 100% certain, and hence there is no need to reflect on how to weigh different uncertain and/or risky outcomes against each other.[135]

[134] Reasoning about risks and uncertainty is about what could happen even if it never does, whereas reasoning about known facts is about what is actually the case.

[135] When one is dealing with risks and uncertainty, one needs, among other things, to grapple with how to weigh uncertainties and risks against actual benefits. One needs to confront the difficult question of why imposing a risk onto somebody might be wrong, even if things go well in the end and certain kinds of actual harms end up not being realized. These and other difficult questions don't arise if, as is rarely the case, one knows exactly what will happen in different scenarios we might instigate (Hayenhjelm & Wolff: 2012). For some discussion of the acceptability of current driving risks, see (Smids: 2016).

Yet again, in other words, we find that the two different issues differ in striking and non-trivial ways. In one case, difficult questions concerning risks and uncertainty immediately arise, whereas in the other, no such issues are involved. This is another important disanalogy between the ethics of accident-algorithms for self-driving cars and the trolley problem. It is a disanalogy that exposes a categorical difference between these two different subjects.

8.7. Concluding Discussion

We have isolated a number of important differences between the ethics of accident-algorithms for self-driving cars and the trolley problem. These all center around three main areas of disanalogy: with respect to the overall decision-situation and its features, with respect to the role of moral and legal responsibility, and with respect to the epistemic situation of the decision-makers. The various points we have made can be summarized and shown with the help of the following table. We here number the main areas of disanalogy as 1 through 3, and sub-divide 1 (viz. the disanalogous features of the basic decision-situations) into the three sub-disanalogies 1a-1c:

	Accident-algorithms for self-driving cars:	**Trolley Problem:**
1a: Decision faced by:	*Groups of individuals/multiple stakeholders*	*One single individual*
1b: Time-perspective:	*Prospective decision/contingency planning*	*Immediate/"here and now"*
1c: Numbers of considerations/situanational features that may be taken into account:	*Unlimited; unrestricted*	*Restricted to a small number of considerations; everything else bracketed*
2: Responsibility, moral and legal:	*Both need to be taken into account*	*Both set aside; not taken into account*
3: Modality of knowledge, or epistemic situation:	*A mix of risk-estimation and decision-making under uncertainty*	*Facts are stipulated to be both certain and known*

We started by asking just how similar, or dissimilar, the trolley problem and the issue of how self-driving cars ought to be programmed are. Return now briefly to the question of whether the literature on the trolley problem is a good, or perhaps the best, place to turn to for input for the ethics of accident-algorithms for self-driving cars. We can now argue as follows.

On the one hand, the key issues we have isolated as being of great importance for the ethics of accident-algorithms for self-driving cars are typically not discussed in the main literature on the trolley problem. For example, this literature is not about the risks or the legal and moral responsibilities we face in traffic. On the other hand, the main issues that the literature on the trolley problem does engage directly with have to do with rather different things than those we have flagged as being most pressing for the ethics of accident-algorithms for self-driving cars. As we noted above, this literature discusses things such as: the ethical differences between positive and negative duties and killing and letting die, and psychological and neuro-scientific theories about how different types of moral judgments are generated by our minds and brains. Taking these considerations together, we think it is clear that the literature on the trolley problem is not the best, nor perhaps even a particularly good, place to turn to for source materials and precedents directly useful for the ethics of accident-algorithms for self-driving cars.

Return next to the positive aim of the chapter, namely, to isolate and identify key issues that the ethics of accident-algorithms for self-driving cars needs to deal with. Based on what we have argued in the previous sections – as summarized in the table above – we wish to draw the following broad conclusions about the general ethical issues that are raised by the question of how to program self-driving cars to respond to accident-scenarios. What we are facing here are complex and difficult ethical issues relating to, among other things, the following:

(i) decision-making faced by groups and/or multiple stake-holders;
(ii) morally loaded prospective decision-making and/or contingency planning;
(iii) open-ended ethical reasoning taking wide ranges of considerations into account
(iv) ethical reasoning concerned with both backward-looking and forward-looking moral and legal responsibility
(v) ethical reasoning about risks and/or decisions under uncertainty.

We add the qualifier "among other things" here in order to make it clear that we are not of the opinion that these are the only general topics that are relevant for the ethics of accident-algorithms for self-driving cars. Rather, it is our view that these are among the general topics that are most relevant for this specific issue, but that there are certainly also other general topics that are highly relevant as we approach this ethical problem in a systematic way. Most importantly, we need to identify the ethical values, considerations, and principles that are best suited to be brought to bear on this pressing ethical issue. And we need to think about how to specify and adapt those values, considerations, and principles to the particular problem of how self-driving cars ought to be programmed to react to accident-scenarios. In other words, there is a lot of work to do here.[136]

[136] For their helpful comments, we are grateful to Hanno Sauer, Noah Goodall, John Danaher, Andreas Spahn, the participants of Eindhoven University of Technology's Philosophy and Ethics Workshop, and two anonymous reviewers for this journal. We are also grateful to Theo Hofman from our university's mechanical engineering department for his help with technical details regarding automated driving.

9 Conclusion

9.1. Conclusions to the research questions

The overarching research question of this thesis is: how can persuasive technologies be designed and used in an ethically responsible way? In chapter 2, I have given a summary treatment of most of the ethical issues concerning design and use of persuasive technologies. This treatment extends the existing literature in two main ways. First, I disambiguate the concept of the 'outcome of persuasion' between user behaviour on the one hand, and the way relevant values are affected by that behaviour on the other hand. As a result, more different routes via which persuasive technologies can have a positive or negative impact become visible. Consequently, our analytical toolbox is expanded, enabling us to better predict and explain the effects of persuasive technology. Second, while existing treatments focus merely on 'designers', I add 'deployers', that is, the party that 'orders' the design of the PT and attempts to get the users use the PT. This addition enables us to see that designers and deployers have different role-responsibilities, the fulfillment of which requires their close cooperation.

I will now summarize the main findings with regard to the four main research questions identified in the introduction. First, the question answered in chapters three and four is under what conditions users of persuasive technology exercise substantial control over their attitudes (and other mental states) and behaviours, such that there is 'voluntary change'. In chapter three I develop three conditions that must be satisfied for non-argumentative means of persuasion to fit rational persuasion. Based on that account, I contrast rational persuasion with various types of nudging, and conclude that governments have reason to combine rational persuasion and nudging in their efforts to influence citizen behaviour. Ideally, governments convince citizens of the desirability of changing their behaviour. In that way, by employing rational persuasion, citizens engage their deliberative capacities and make up their own minds, and consequently, governments do not have to paternalistically decide for citizens how they should behave and to nudge them subsequently. Instead, citizens convinced that they should eat healthier or consume less energy, might very well be willing to accept nudges that support them.

Building on my account of rational persuasion, I propose three conditions that must be satisfied such that users of persuasive technology exercise substantial control over their attitudes and behaviour:

(i) The PT's communication should *provide information that supports* the user in determining whether the target behaviour is practically rational for her.
(ii) PTs should *not subvert or undermine* the user's practical rationality.
(iii) The PT should also grant the user the *freedom to act on the outcome* of her deliberation.

This answer to the first research question is crucial for answering the second, which concerns how to best characterize persuasive technology (Chapter 4). By way of giving this positive specification of what must be the case for substantial control, we are able to say more than that persuasive technology should not involve coercion or deception, but instead relies on voluntary change. Together with the key idea that persuasion is a form of communication, this specification results in the following definition of persuasive technology:

Persuasive technologies are technologies that are (i) intentionally (ii) designed to change some mental state(s) of the user, most often with the ultimate aim of behaviour change. They do so (iii) by communicating (iv) in a way that grants users substantial control over their mental states and behaviour.

By applying this redefinition, the distinctions between persuasive technology on the one hand, and manipulative technology, coercive technology, and 'limiting' technology follow in a straightforward manner. This is a helpful result for performing ethical analysis of persuasive technology.

The third main question of the research of this thesis is how to morally evaluate the use of social influence strategies in the design of persuasive technology. More specifically, I focus on 'similarity influence' (i.e. influence based on ways in which designers have designed the persuasive technology to appear similar to the user, such as the PT mimicking the user's head movement).

I argue that quite often similarity influence will work in a way that leaves users less than substantial control over what they think and do. I extensively discuss mimicry, which normally happens unintentionally and unnoticed, involving automatic psychological processes. However, when designers incorporate similarity influence into a persuasive technology, they intentionally enhance user trust and cooperation outside user's awareness. In this way, the user's practical rationality is undermined and the charge of manipulation is justified. By means of extending some key ideas of Habermas' ethics of communication, I propose three design guidelines that together should safeguard substantial user control in case of PT involving similarity-based influence. While Chapter 5 may deliver some useful ideas and tentative conclusions, what becomes most obvious is the need for further study on this topic. I will say a bit more about this below.

The fourth main research question concerns the conditions of acceptable risk imposition, inspired by the fact that an important class of persuasive technology is designed for risk-reduction. Scanlon's contractualism is found to provide an attractive middle ground between consequentialist and rights-based or Kantian approaches. Despite criticisms to the contrary, Scanlon manages to avoid *inter*-personal aggregation, which is the root cause of the problems of consequentialist approaches. His contractualism does not permit justifying serious risk to one individual on the basis of small benefits to many others. At the same time, Scanlon seems to successfully avoid the problem of paralysis that besets Kantian or rights-based approaches. However, his commitment to *intra*-personal aggregation still leads to rather stringent restrictions on acceptable risk imposition, that many still find unreasonable. In any case, Scanlon's restriction to intrapersonal aggregation as the only way in which risk burdens and benefits may be compared and weighed leads to interesting implications. Most importantly, our society should not focus so much on the acceptability of single risk-generating actions or even practices. Instead, we should focus on the cumulative risk-exposure during lifetime for different individuals, and ensure that these are sufficiently equal (see also below).

Building on the findings in Chapter 6, in Chapter 7 I argue for a conclusive moral case in favor of mandatory use by drivers of both warning and limiting versions of Intelligent Speed Adaptation. Current driving risks are unacceptable due to their magnitude, their distribution, and the existence of many (cost-effective) ways to significantly reduce them. Intelligent Speed Adaptation, then, is justified by its large potential to reduce these risks. Upon closer examination, I find all possible objections to be unconvincing. The technology that is the topic of Chapter 8, finally, is located higher on the control continuum than persuasive technology. We (Sven Nyholm and the present author) ask whether the problem of programming accident algorithms for self-driving cars is an applied trolley problem. On the basis of three major disanalogies, we argue that it is not. One of these disanalogies concerns the epistemic situation of those that have to decide whom to let live and whom to kill. In the trolley problem, the deciding person has full knowledge of outcomes that are certain. Programming accident algorithms, on the contrary, involves uncertainty due to probabilistic outcomes. We do not fully know the probabilities and we clearly need a sound framework of morally acceptable risk in order to program the right choices.

9.2. Avenues for further study

The research of this thesis reveals several interesting and important avenues for further study, two of which I will now briefly discuss. First, it has become clear that the design of persuasive technology as a kind of social actor, often incorporating similarity-based influence, raises a host of intricate and philosophically challenging questions. The prevalence of artificial or digital social agents with many added persuasive features is increasing quickly, which makes it urgent to address these issues. Often, it requires subtle distinctions to give an ethical evaluation of the use of social influence between humans (cf. Buss: 2005). It is useful to consider what is different in an ethically relevant manner when we transfer these influences from their 'natural' context of human - human interaction to the context of persuasive technology - human interaction. I propose that, as is the case with my detailed discussion of mimicry, a (Habermassian) ethics of communication is a valuable source of insight for further pursuing this question.

The second topic for further study concerns the conditions of acceptable risk-imposition (Chapter 6). The discussion of these conditions leads us to deep and fundamental questions regarding our entitlements or rights to act in ways that generate risk to others *and* to be free from risks at serious harm. For Scanlon's contractualism, the question is whether individuals already have an entitlement to act in ways that create risk, which can serve as input of contractualist reasoning about which principles no one can reasonably reject. Alternatively, such entitlement could instead be regarded as the outcome of contractualist reasoning. At present, it seems unclear whether Scanlon's contractualism provides a principled way to answer this question.

For rights-based theories, a similar question arises as to whether we have two independent, *sui generis*, rights to agency and to be free from serious risks. If we do not, which of these rights is the more fundamental? And can one be derived from the other? The answers will depend on the justification given for the rights we have, for example along the lines of the natural rights approach, Gewirth's agency based justification (cf. Gewirth: 1998), or some other justification. In any case, my hypothesis is that once we grant individuals a right to risk-generating agency from the outset, we inevitably end up with unequal levels of mutual risk-imposition. Then, as soon as some individual is exposed to a higher risk level than that she imposes on others, this extra risk is no longer intra-personally beneficial to her (see Chapter 6 above). Consequently, a justification of imposing risks on others on the basis of intra-personal aggregation alone becomes impossible.

While a right to agency is meaningless without the acknowledgement that such a right entails risk, I propose that this risk should be limited to an absolute minimum. Subsequently, we should search for other sources of justification for higher levels of mutual risk-imposition, in ways that do not involve problematic inter-personal aggregation of risks and benefits. As long as we lack adequate answers to this set of fundamental questions, we do well to adhere to the plurality of principles for guiding acceptable risk mentioned in Chapter 6.

9.3. Policy recommendations

Finally, the present research gives reason for a number of recommendations regarding the design and use of persuasive technology. A first recommendation concerns what researchers can do to help safeguard substantial control of users over their mental states and behaviour. Determining the extent to which persuasive strategies and methods employed in a persuasive technology grant users such control requires detailed knowledge of the underlying psychological processes. Given that researchers studying and designing persuasive technologies have such expert knowledge, it would be very helpful for ethical reflection if they share that knowledge with ethicists. One way to do so is to develop a practice in which researchers and designers specify the means of persuasion they have employed, together with a description of the type of psychological processes by which these means exert their influence on users. Preferably, they describe characteristics such as the likelihood that these processes bypass or subvert user reflection, the chances for motivated users to make up their own mind regarding what to believe and do, the amount of relevant information for doing so given by the persuasive technology, and perhaps still other characteristics. In this way, each persuasive strategy or prototype design should be accompanied by a kind of information leaflet, similar to a patient package insert. Such a practice would be of great help for a responsible design of persuasive technology.

Second, given that most ethical issues with regard to persuasive technology center around the user, it is recommended to involve the user in the design. Developing persuasive technologies often leads to value conflicts that involve users in the first place. Given their right to autonomy, they should be involved in the design in order to co-determine how to deal with these conflicts. During last decades, several design approaches have been developed that involve the user in the design, or at least pay close attention to user values. In Chapter 2, I already mentioned two of them, Value Sensitive Design, and Participatory Design, but still others exist.

Regarding the long-term consequences of widespread use of persuasive technology, third, our governments would do well to stimulate a public debate and to fund research. This is a task for governments, because individual parties have no role-obligation and often no incentive to address this issue. As discussed in Chapter 2, there is a real question of what effects long-term and extensive use of persuasive technologies, such as health-apps, will have on our distinct human capacities of deliberation, decision-making, will-power, and the like. The worry is that these capacities weaken in case we rely too much on supportive persuasive technologies. In this way, we risk losing what makes us distinctly human. Moreover, in the long term this process would be self-defeating. The first stage of the development of any persuasive technology is to determine its ends, its final values. Doing so requires reflection and debate, often public debate, for which the very same distinct human capacities are indispensable.

Finally, and closely related, governments should not give up on education as a means to address societal problems, expecting instead to solve all problems with persuasive technology, nudging, and the similar ways to influence citizens. In fact, education is still *the* path to the most thorough, sustainable, effective, and productive changes in attitudes and behaviour. Most fundamentally, without education we would not develop the capacity to determine which attitudes and behaviours are most worthwhile in the first place.

Bibliography

Aarts, E. and Wichert, R. (2009), 'Ambient intelligence', in: Bullinger, H.-J. (ed.), *Technology Guide*, Springer Berlin Heidelberg: 244–249.

Aarts, L. and van Schagen, I. (2006) 'Driving speed and the risk of road crashes: A review', in: *Accident Analysis & Prevention* 38(2): 215–224.

Achenbach, J. (2015), 'Driverless cars are colliding with the creepy Trolley Problem', in: *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/

Ajzen, I. (2012) 'Martin Fishbein's Legacy The Reasoned Action Approach', in: *The ANNALS of the American Academy of Political and Social Science* 640(1), 11–27. https://doi.org/10.1177/0002716211423363

Ajzen, I. and Fishbein, M. (2000), 'Attitudes and the Attitude-Behavior Relation: Reasoned and Automatic Processes. in: *European Review of Social Psychology* 11 (1): 1–33.

Altham, J. E. J. (1983) 'Ethics of Risk', in: *Proceedings of the Aristotelian Society* 84: 15–29.

Anderson, J. H. (2010) 'Review of Richard Thaler and Cass Sunstein: Nudge: Improving Decisions about Health, Wealth, and Happiness', in: *Economics and Philosophy* 26 (3): 369–376.

Anderson, J.H. and Kamphorst, B.A. (2014), 'Ethics of e-coaching: Implications of employing pervasive computing to promote healthy and sustainable lifestyles', in: *2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*: 351–356.

Anderson, J.H and Kamphorst, B. A. (2015) 'Should Uplifting Music and Smart Phone Apps Count as Willpower Doping? The Extended Will and the Ethics of Enhanced Motivation', in: *AJOB Neuroscience* 6 (1): 35–37.

Anderson, S. (2011), 'Coercion', in: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011). Retrieved from http://plato.stanford.edu/archives/win2011/entries/coercion/

Aristotle. (1991). *The art of rhetoric*. (H. Lawson-Tancred, Trans.), Penguin, London.

Ashford, E. (2003) 'The Demandingness of Scanlon's Contractualism', in: *Ethics* 113 (2): 273–302.

Ashford, E. and Mulgan, T. (2012), 'Contractualism', in: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012). Retrieved from http://plato.stanford.edu/archives/fall2012/entries/contractualism/

Azar, K. M. J., Lesser, L. I., Laing, B. Y., Stephens, J., Aurora, M. S., Burke, L. E. and Palaniappan, L. P. (2013) 'Mobile Applications for Weight Management', in: *American Journal of Preventive Medicine* 45 (5): 583–589.

Bailenson, J. N., Garland, P., Iyengar, S. and Yee, N. (2006) 'Transformed Facial Similarity as a Political Cue: A Preliminary Investigation', in: *Political Psychology* 27 (3): 373–385.

Bailenson, J. N. and Yee, N. (2005) 'Digital Chameleons Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments', in: *Psychological Science* 16 (10): 814–819.

Bang, M. and Ragnemalm, E. L. (eds.) (2012) Persuasive Technology. Design for Health and Safety (Vol. 7284), Springer, Berlin/Heidelberg.

Bargh, J. A. and Chartrand, T. L. (1999) 'The unbearable automaticity of being', in: *American Psychologist;American Psychologist* 54 (7): 462–479.

Barkenbus, J. N. (2010) 'Eco-driving: An overlooked climate change initiative', in: *Energy Policy* 38 (2): 762–769.

Baron, M. (2003) 'Manipulativeness', in: *Proceedings and Addresses of the American Philosophical Association* 77 (2): 37–54.

Barral, O., Aranyi, G., Kouider, S., Lindsay, A., Prins, H., Ahmed, I., ... Cavazza, M. (2014), 'Covert Persuasive Technologies: Bringing Subliminal Cues to Human-Computer Interaction', in: *Persuasive Technology*, Springer, Cham: 1–12.

Barth, M. and Boriboonsomsin, K. (2009) 'Energy and emissions impacts of a freeway-based dynamic eco-driving system', in: *Transportation Research Part D: Transport and Environment* 14 (6): 400–410.

Baumann, H. and Döring, S. (2011), 'Emotion-Oriented Systems and the Autonomy of Persons', in: Cowie, R., Pelachaud, C. and P. Petta (eds.), *Emotion-Oriented Systems,* Springer, Berlin/Heidelberg: 735–752.

Beauchamp, T. L. and Childress, J. F. (2009) Principles of Biomedical Ethics, Oxford University Press, New York/Oxford.

Benn, S. I. (1967) 'Freedom and Persuasion', in: *Australasian Journal of Philosophy* 45 (3): 259–275.

Berdichevsky, D. and Neuenschwander, E. (1999) 'Toward an ethics of persuasive technology', in: *Commun. ACM* 42 (5): 51–58.

Berends, E. M. (2009) De invloed van automassa op het letsel risico bij botsingen tussen twee personenauto's: een kwantitatieve analyse. [The impact of car-mass on injury risk from crashes between two person vehicles: a quantitative analysis], SWOV, Den Haag.

Berkovsky, S., Freyne, J. and Oinas-Kukkonen, H. (eds.) (2012) 'Influencing Individually: Fusing Personalization and Persuasion', in: *ACM Trans. Interact. Intell. Syst.* 2 (2): 9:1–9:8.

Berlo, D. (1960) Process of Communication: An Introduction to Theory and Practice, Harcourt School, New York.

Bhatnagar, N., Sinha, A., Samdaria, N., Gupta, A., Batra, S., Bhardwaj, M. and Thies, W. (2012) 'Biometric Monitoring as a Persuasive Technology: Ensuring Patients Visit Health Centers in India's Slums', in: In Bang, M. and Ragnemalm, E. L. (eds.), *Persuasive Technology. Design for Health and Safety*, Springer, Berlin/Heidelberg: 169–180.

Bickmore, T. W., Schulman, D. and Sidner, C. (2013) 'Automated interventions for multiple health behaviors using conversational agents', in: *Patient Education and Counseling* 92 (2): 142–148.

Blair, J. (2012) 'Argumentation as Rational Persuasion', in: *Argumentation* 26 (1): 71–81.

Blumenthal-Barby, J. S. (2014), 'A Framework for Assessing the Moral Status of "Manipulation"', in: Coons, C. and Weber, M. (eds.), *Manipulation*, Oxford University Press, Oxford: 121–134.

Bohman, J. and Rehg, W. (2014), 'Jürgen Habermas', in: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2014), Retrieved from https://plato.stanford.edu/archives/fall2014/entries/habermas/

Bonnefon, J.-F., Shariff, A. and Rahwan, I. (2015) 'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?', in: *ArXiv:1510.03346 [Cs]*. Retrieved from http://arxiv.org/abs/1510.03346

Bovens, L. (2009), 'The Ethics of Nudge', in: Grüne-Yanoff, T. and Hansson, S. O. (eds.), *Preference Change* Springer Netherlands, Dordrecht: 207–219.

Breton, E. R., Fuemmeler, B. F. and Abroms, L. C. (2011), 'Weight loss—there is an app for that! But does it adhere to evidence-informed practices?', in: *Translational Behavioral Medicine* 1 (4): 523–529.

Briñol, P. and Petty, R. E. (2006) 'Fundamental Processes Leading to Attitude Change: Implications for Cancer Prevention Communications', in: *Journal of Communication* 56: S81–S104.

Brownsword, R. (2005) 'Code, control, and choice: why East is East and West is West', in: *Legal Studies* 25 (1): 1–21.

Brownsword, R. and Goodwin, M. (2012). Law and the technologies of the twenty-first century: text and materials, Cambridge University Press, Cambridge, UK/New York.

Burnell, P. and Reeve, A. (1984) 'Persuasion as a Political Concept.', in: *British Journal of Political Science* 14 (4): 393–410.

Buss, S. (2005) 'Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints', in: *Ethics* 115 (2): 195–235.

Carsten, O. (2012) 'Is intelligent speed adaptation ready for deployment?', in: *Accident Analysis and Prevention* 48: 1–3.

Carsten, O. M. J. and Tate, F. N. (2005) 'Intelligent speed adaptation: accident savings and cost–benefit analysis', in: *Accident Analysis & Prevention* 37 (3): 407–416.

Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. F. (eds.) (2000) Embodied Conversational Agents, The MIT Press, Cambridge/Mass.

Cave, E. (2007) 'What's Wrong with Motive Manipulation?', in: *Ethical Theory and Moral Practice* 10 (2): 129–144.

Chaiken, S. and Maheswaran, D. (1994) 'Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment', in: *Journal of Personality and Social Psychology* 66 (3): 460–473.

Chartrand, T. L. and Bargh, J. A. (1999) 'The chameleon effect: The perception–behavior link and social interaction', in: *Journal of Personality and Social Psychology* 76 (6): 893–910.

Chartrand, T. L., Maddux, W. W. and Lakin, J. L. (2006), 'Beyond the Perception-Behavior Link: The Ubiquitous Utility and Motivational Moderators of Nonconscious Mimicry', in: Hassin, R. R., Uleman, J. S. and Bargh, J. A. (eds.), *The New Unconscious,* Oxford University Press, Oxford: 334–361.

Chittaro, L. (2012) 'Passengers' Safety in Aircraft Evacuations: Employing Serious Games to Educate and Persuade', in: *Persuasive Technology. Design for Health and Safety,* Springer, Berlin/Heidelberg: 215–226.

Christman, J. (2011) The Politics of Persons: Individual Autonomy and Sociohistorical Selves (Reissue edition). Cambridge University Press, Cambridge.

----- (2015), 'Autonomy in Moral and Political Philosophy', in:. Zalta, E. N (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2015). Retrieved from http://plato.stanford.edu/archives/spr2015/entries/autonomy-moral/

Cialdini, R. B. (2006) Influence: The Psychology of Persuasion (Revised edition), Harper Business, New York NY.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K. and Winter, P. L. (2006) 'Managing social norms for persuasive impact', in: *Social Influence* 1 (1): 3–15.

Cohen, S. (2013) 'Nudging and Informed Consent', in: *The American Journal of Bioethics* 13 (6): 3–11.

Collingridge, D. (1981) The Social Control of Technology, Palgrave Macmillan, New York.

Compen, N., Spahn, A. and Ham, J. (2015), 'Sustainability coaches. A better environment starts with your coach', in: Kool, L, Timmer, J. and van Est, R. (eds.), *Sincere Support. The Rise of the E-coach.* Rathenau Instituut, Den Haag. Retrieved from https://www.rathenau.nl/nl/publicatie/sincere-support-rise-e-coach

Cranor, C. F. (2007), 'Towards a non-consequentialist approach to acceptable risks', in: Lewens, T. (ed.), *Philosophical Perspectives on Risk,* Routledge, London/New York:.

Davis, J. (2010), 'Generating Directions for Persuasive Technology Design with the Inspiration Card Workshop', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), *Persuasive Technology*, Springer, Berlin/Heidelberg: 262–273.

de Sousa, R. (2014), 'Emotion', in: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Retrieved from https://plato.stanford.edu/archives/spr2014/entries/emotion/

DeBruine, L. M. (2002) 'Facial resemblance enhances trust', in: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269 (1498): 1307–1312.

DeCew, J. (2015) 'Privacy', in: Zalta, E.N. (ed), *The Stanford Encyclopedia of Philosophy* (Spring 2015). Retrieved from http://plato.stanford.edu/archives/spr2015/entries/privacy/

Dijksterhuis, A., Aarts, H. and Smith, P. K. (2006) 'The Power of the Subliminal: On Subliminal Persuasion and Other Potential Applications', in: Hassin, R. R., Uleman, J. S. and Bargh, J. A. (eds.), *The New Unconscious,* Oxford University Press, Oxford: 77–106.

Dijkstra, A. (2006), 'Technology Adds New Principles to Persuasive Psychology: Evidence from Health Education', in: IJsselsteijn, W. A., de Kort, Y. A. W., Midden, C., Eggen, B. and van den Hoven, E. (eds.), *Persuasive Technology* Springer, Berlin/Heidelberg: 16–26.

Doctorow, C. (2015), 'The problem with self-driving cars: who controls the code?' *The Guardian*. Retrieved from http://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code

Doorn, N. (2011) Moral Responsibility in R&D Networks. Technische Universiteit Delft, Delft.

Dworkin, G. (1988) The theory and practice of autonomy, Cambridge University Press, Cambridge/New York.

----- (2013) 'Lying and nudging', in: *Journal of Medical Ethics* 39 (8): 496–497.

----- (2016), 'Paternalism', in: Zalta, E.N. (ed),*The Stanford Encyclopedia of Philosophy* (Summer 2016). Retrieved from http://plato.stanford.edu/archives/sum2016/entries/paternalism/

Elvik, R. (2009). The Handbook of Road Safety Measures (2nd Revised edition edition), Emerald Group Publishing Limited, Bingley UK.

Elvik, R. (2014) 'Speed and Road Safety - New Models', *TØI Report*, (1296/2014). Retrieved from http://trid.trb.org/view.aspx?id=1318116

European Road Safety Observatory (2012a) 'Annual Statistical Report 2012', European Commission, DG Mobility & Transport. Retrieved from http://ec.europa.eu/transport/road_safety/pdf/statistics/dacota/dacota-3.5-asr-2012.pdf

European Road Safety Observatory (2012b) 'Traffic Safety Basic Facts 2012: Main Figures', European Commission, DG Mobility & Transport. Retrieved from http://ec.europa.eu/transport/road_safety/pdf/statistics/dacota/bfs2012_dacota-trl-main_figures.pdf

European Road Safety Observatory. (2012c). 'Traffic Safety Basic Facts 2012: Single vehicle accidents', European Commission, DG Mobility & Transport. Retrieved from http://ec.europa.eu/transport/road_safety/pdf/statistics/dacota/bfs2012-dacota-ntua-single_vehicle_accident.pdf

Evans, J. S. B. T. and Stanovich, K. E. (2013) 'Dual-Process Theories of Higher Cognition Advancing the Debate', in *Perspectives on Psychological Science* 8 (3): 223–241.

Evans, L. (2001) 'Causal influence of car mass and size on driver fatality risk', in: *American Journal of Public Health* 91 (7): 1076–1081.

----- (2008) 'Death in Traffic: Why Are the Ethical Issues Ignored?', in: *Studies in Ethics, Law, and Technology* 2 (1):https://doi.org/10.2202/1941-6008.1014

Eyal, N. (2012), 'Informed Consent', in: Zalta, E.N. (ed), *The Stanford Encyclopedia of Philosophy* (Fall 2012). Retrieved from http://plato.stanford.edu/archives/fall2012/entries/informed-consent/

Faden, R. R. and Beauchamp, T. L. (1986) A History and Theory of Informed Consent, Oxford University Press, New York.

Fahlquist, J. N. (2009) 'Saving lives in road traffic—ethical aspects', in: *Journal of Public Health* 17 (6): 385–394.

Feinberg, J. (1987) Harm to Others. The moral limits of the criminal law (Vol 1), Oxford University Press, New York/Oxford.

----- (1989) Harm to Self. The Moral Limits of the Criminal Law (Vol 3), Oxford University Press, New York/Oxford.

Fiske, S. and Taylor, S. (1984) Social Cognition, Longman Higher Education.

Fogg, B. (2003) Persuasive technology : using computers to change what we think and do, Morgan Kaufmann Publishers, Amsterdam/Boston.

Foot, P. (1967) 'The problem of abortion and the doctrine of double effect', in: *The Oxford Review* 5.

Franssen, M., Lokhorst, G.-J. and van de Poel, I. (2015), 'Philosophy of Technology', in: Zalta, E.N. (ed),*The Stanford Encyclopedia of Philosophy* (Fall 2015). Retrieved from https://plato.stanford.edu/archives/fall2015/entriesechnology/

Fried, B. H. (2012) 'Can Contractualism Save Us from Aggregation?', in: *The Journal of Ethics* 16 (1): 39–66.

Fried, C. (1970) An Anatomy of Values: Problems of Personal and Social Choice, Harvard University Press, Cambridge Mass.

Friedman, B., Jr, P. H. K., Borning, A. and Huldtgren, A. (2013), ‘Value Sensitive Design and Information Systems’, in: Doorn, N. Schuurbiers, D. van de Poel, I. and Gorman, M. E. (eds.), *Early engagement and new technologies: Opening up the laboratory*, Springer Netherlands: 55–95

Friedman, B. and Kahn, P. H., Jr. (2003), ‘Human Values, Ethics, and Design’, in: Jacko, J. A. and Sears, A. (eds.), *The Human-computer Interaction Handbook*, L. Erlbaum Associates Inc, Hillsdale NJ: 1177–1201.

Fung, B. (2017) ‘The driver who died in a Tesla crash using Autopilot ignored at least 7 safety warnings’, in: *Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2017/06/20/the-driver-who-died-in-a-tesla-crash-using-autopilot-ignored-7-safety-warnings/

Gewirth, A. (1998) The Community of Rights, University Of Chicago Press, Chicago.

Gogoll, J. and Müller, J. F. (2017) ‘Autonomous Cars: In Favor of a Mandatory Ethics Setting’, in: *Science and Engineering Ethics* 23 (3): 681-700.

Goldman, A. I. (1999) Knowledge in a Social World, Clarendon Press, Oxford/ New York.

Goodall, N. (2014a) ‘Ethical Decision Making During Automated Vehicle Crashes’, in: *Transportation Research Record: Journal of the Transportation Research Board* 2424: 58–65.

Goodall, N. J. (2014b), ‘Machine Ethics and Automated Vehicles’, in: Meyer, G. and Beiker, S. (eds.), *Road Vehicle Automation,* Springer International Publishing: 93–102.

Gorin, M. (2014) ‘Do Manipulators Always Threaten Rationality?’ *American Philosophical Quarterly* 51 (1): 51-61.

Green, M. (2009), ‘Speech Acts’, in: Zalta, E.N. (ed), *The Stanford Encyclopedia of Philosophy* (Spring 2009). Retrieved from http://plato.stanford.edu/archives/spr2009/entries/speech-acts/

Greene, J. (2013) Moral Tribes: Emotion, Reason, and the Gap Between Us and Them, Penguin Books.

Grüne-Yanoff, T. (2012) ‘Old wine in new casks: libertarian paternalism still violates liberal principles’, in: *Social Choice and Welfare* 38 (4): 635–645.

----- (2016) ‘Why Behavioural Policy Needs Mechanistic Evidence’, in: *Economics and Philosophy* 32 (3): 463–483.

Grüne-Yanoff, T. and Hertwig, R. (2016) ‘Nudge Versus Boost: How Coherent are Policy and Theory?’, in: *Minds and Machines* 26 (1–2): 149–183.

Habermas, J. (1985) The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society. [T. McCarthy, Trans.], Beacon Press, Boston.

Habermas, J. (1998) On the Pragmatics of Communication, The MIT Press, Cambridge Mass.

Hale, J. and Hamilton, A. F. D. C. (forthcoming) 'Testing the relationship between mimicry, trust and rapport in virtual reality conversations', in: *Scientific Reports* 6. https://doi.org/10.1038/srep35295

Ham, J. and Midden, C. (2010), 'Ambient Persuasive Technology Needs Little Cognitive Effort: The Differential Effects of Cognitive Load on Lighting Feedback versus Factual Feedback', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), *Persuasive Technology*, Springer, Berlin/Heidelberg: 262–273.

Ham, J. and Spahn, A. (2015) 'Shall I Show You Some Other Shirts Too? The Psychology and Ethics of Persuasive Robots', in: *A Construction Manual for Robots' Ethical Systems*, Springer, Cham: 63–81.

Hamari, J., Koivisto, J. and Pakkanen, T. (2014), 'Do Persuasive Technologies Persuade? - A Review of Empirical Studies', in: Spagnolli, A. Chittaro, L. and. Gamberini, L (eds.), *Persuasive Technology,* Springer International Publishing: 118–136.

Hansson, S. O. (2012), 'A Panorama of the Philosophy of Risk', in: Roeser, S. Hillerbrand, R. Sandin, P. and Peterson, M. (eds.), *Handbook of Risk Theory,* Springer Netherlands: 27–54.

----- (2013). The Ethics of Risk: Ethical Analysis in an Uncertain World, Palgrave Macmillan Houndsmills, Basingstoke, Hampshire; New York, NY:.

----- (2003) 'Ethical Criteria of Risk Acceptance', in: *Erkenntnis* 59 (3): 291–309.

Harsanyi, J. C. (1953) 'Cardinal Utility in Welfare Economics and in the Theory of Risk-taking', in: *Journal of Political Economy* 61 (5): 434–435.

----- (1982), 'Morality and the Theory of Rational Behavior', in: Sen, A. and Williams, B. (eds.), *Utilitarianism and Beyond*, Cambridge University Press, Cambridge: 39-62.

Hausman, D. M. and Welch, B. (2010) 'Debate: To Nudge or Not to Nudge*', in: *Journal of Political Philosophy* 18 (1): 123–136.

Hayenhjelm, D. M. (2012), 'What Is a Fair Distribution of Risk?' in: Roeser, S. Hillerbrand, R. Sandin, P. and Peterson, M. (eds.), *Handbook of Risk Theory,* Springer Netherlands: 27–54.

Hayenhjelm, M. and Wolff, J. (2012) 'The Moral Problem of Risk Impositions: A Survey of the Literature', in: *European Journal of Philosophy* 20, E26–E51.

Hevelke, A. and Nida-Rümelin, J. (2014) 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis', in: *Science and Engineering Ethics* 21 (3): 619–630.

Hogg, M. A., Hohmann, Z. P. and Rivera, J. E. (2009) 'Teaching and Learning Guide for: Why Do People Join Groups? Three Motivational Accounts from Social Psychology', in: *Social and Personality Psychology Compass* 3 (1): 111–117.

Holm, S. and Ploug, T. (2013) '"Nudging" and Informed Consent Revisited: Why "Nudging" Fails in the Clinical Context', *The American Journal of Bioethics* 13 (6): 29–31.

Holzwarth, M., Janiszewski, C. and Neumann, M. M. (2006) 'The Influence of Avatars on Online Consumer Shopping Behavior', in: *Journal of Marketing* 70 (4): 19–36.

Honda Worldwide (2008) 'Honda Develops Ecological Drive Assist System for Enhanced Real World Fuel Economy'. Retrieved 22 January 2016, from http://world.honda.com/news/2008/4081120Ecological-Drive-Assist-System/

Hooker, B. (2002) 'Contractualism, spare wheel, aggregation', in: *Critical Review of International Social and Political Philosophy* 5 (2): 53–76.

Horcajo, J., Petty, R. E. and Briñol, P. (2010) 'The effects of majority versus minority source status on persuasion: A self-validation analysis', in: *Journal of Personality and Social Psychology* 99 (3): 498–512.

Huckvale, K., Prieto, J. T., Tilney, M., Benghozi, P.-J. and Car, J. (2015) 'Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment', in: *BMC Medicine* 13: 214.

Husak, D. (2004) 'Vehicles and Crashes: Why is this Moral Issue Overlooked?', in: *Social Theory and Practice* 30 (3): 351–370.

IJsselsteijn, W. A., de Kort, Y. A. W., Midden, C., Eggen, B., and van den Hoven, E. (eds.) (2006) *Persuasive Technology,* (Vol. 3962). Springer, Berlin/Heidelberg.

IJsselsteijn, W., Kort, Y., Midden, C., Eggen, B., and Hoven, E. (2006). Persuasive Technology for Human Well-Being: Setting the Scene. in: IJsselsteijn, W. A., de Kort, Y. A. W., Midden, C., Eggen, B. and van den Hoven, E. (eds.), *Persuasive Technology* Springer, Berlin/Heidelberg: 1–5.

James, A. (2004) 'Rights and Circularity in Scanlon's Contractualism', in: *Journal of Moral Philosophy* 1 (3): 367–374.

----- (2012) 'Contractualism's (Not So) Slippery Slope', in: *Legal Theory* 18 (Special Issue 03): 263–292.

Jamson, S. (2006) 'Would those who need ISA, use it? Investigating the relationship between drivers' speed choice and their use of a voluntary ISA system', in: *Transportation Research Part F: Traffic Psychology and Behaviour* 9 (3): 195–206.

Jamson, S., Chorlton, K. and Carsten, O. (2012) 'Could Intelligent Speed Adaptation make overtaking unsafe?', in: *Accident Analysis & Prevention* 48: 29–36.

Kahneman, D. (2003) 'Maps of Bounded Rationality: Psychology for Behavioral Economics', in: *The American Economic Review* 93 (5): 1449–1475.

Kaipainen, K., Mattila, E., Kinnunen, M.-L. and Korhonen, I. (2010) 'Facilitation of Goal-Setting and Follow-Up in an Internet Intervention for Health and Wellness', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), *Persuasive Technology*, Springer, Berlin/Heidelberg: 238–249.

Kallgren, C. A., Reno, R. R. and Cialdini, R. B. (2000) 'A Focus Theory of Normative Conduct: When Norms Do and Do not Affect Behavior', in: *Personality and Social Psychology Bulletin* 26 (8): 1002–1012.

Kamm, F. M. (2015) The Trolley Problem Mysteries, Oxford University Press, Oxford/New York.

Kampf, R. (2016), 'Long-Term Effects of Computerized Simulations in Protracted Conflicts: The Case of Global Conflicts', in: *Persuasive Technology,* Springer, Cham: 242–247. https://doi.org/10.1007/978-3-319-31510-2_21

Kamphorst, B. and Kalis, A. (2014) 'Why option generation matters for the design of autonomous e-coaching systems', in: *AI & SOCIETY* 30 (1): 77–88.

Kant, I. (1788/2003) Groundwork for the Metaphysics of Morals (A. Zweig, Trans.), Oxford University Press, Oxford/New York.

Kaptein, M. and Eckles, D. (2010)'Selecting Effective Means to Any End: Futures and Ethics of Persuasion Profiling', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), *Persuasive Technology*, Springer, Berlin/Heidelberg: 82–93.

Kaptein, M., Markopoulos, P., de Ruyter, B. and Aarts, E. (2010) 'Persuasion in ambient intelligence', in: *Journal of Ambient Intelligence and Humanized Computing* 1 (1): 43–56.

Kaptein, M., Markopoulos, P., de Ruyter, B. and Aarts, E. (2011) 'Two acts of social intelligence: the effects of mimicry and social praise on the evaluation of an artificial agent', in: *AI & Society* 26 (3): 261–273.

Keating, G. C. (2003) 'Pressing precaution beyond the point of cost-justification', *Vanderbilt Law Review* 56 (3): 651–748.

Kool, L., Timmer, J., and Est, R. van. (2015a), 'E-coaching: from possible to desirable', in: Kool, L. Timmer, J. and. van Est, R (eds.), *Sincere Support. The Rise of the E-coach.* Rathenau Instituut, Den Haag: 9-29.

Kool, L., Timmer, J. and Est, R. van (eds.). (2015b). *Sincere support. The rise of the e-coach.* Rathenau Instituut, Den Haag. Retrieved from https://www.rathenau.nl/nl/publicatie/sincere-support-rise-e-coach

Lai, F., Carsten, O. and Tate, F. (2012) 'How much benefit does Intelligent Speed Adaptation deliver: An analysis of its potential contribution to safety and environment', in: *Accident Analysis & Prevention* 48: 63–72.

Latour, B. (1992) 'Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts', in: Bijker, W. E. and Law, J. (eds.), *Shaping Technology/Building Society: Studies in Sociotechnical Change,* MIT Press, Cambridge Mass: 225–258.

Lehto, T. and Oinas-Kukkonen, H. (2010), 'Persuasive Features in Six Weight Loss Websites: A Qualitative Evaluation', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), Persuasive Technology, Springer, Berlin/Heidelberg: 162–173.

Leighton, J., Bird, G., Orsini, C. and Heyes, C. (2010) 'Social attitudes modulate automatic imitation', in: *Journal of Experimental Social Psychology* 46 (6): 905–910.

Lenman, J. (2008) 'Contractualism and risk imposition', in: *Politics, Philosophy & Economics* 7 (1): 99–122.

Lin, P. (2013) 'The Ethics of Autonomous Cars', in: *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

----- (2015), 'Why Ethics Matters for Autonomous Cars', in: Maurer, M. Gerdes, J. C. Lenz, B. and Winner, H. (eds.), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte,* Springer, Berlin/Heidelberg: 69–8.

Linder, C. (2014) 'Are Persuasive Technologies Really Able to Communicate?: Some Remarks to the Application of Discourse Ethics', in: *International Journal of Technoethics* 5 (1): 44–58.

Lomasky, L. E. (1997) 'Autonomy and Automobility' in: *The Independent Review* 2 (1): 5–28.

Loomis, D., Grosse, Y., Lauby-Secretan, B., Ghissassi, F. E., Bouvard, V., Benbrahim-Tallaa, L., ... Straif, K. (2013) 'The carcinogenicity of outdoor air pollution', in: *The Lancet Oncology* 14 (13): 1262–1263.

MacLean, D. (1986), 'Risk and Consent. Philosophical Issues for Centralized Decisions', in: MacLean, D. (ed.), *Values at Risk,* Rowman and Allanheld, Totowa N.J.

Maddux, W. W., Mullen, E. and Galinsky, A. D. (2008) 'Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates negotiation outcomes', in: *Journal of Experimental Social Psychology* 44 (2): 461–468.

Margolis, E., and Laurence, S. (2014), 'Concepts', in: Zalta, E.N. (ed),*Stanford Encyclopedia of Philosophy* (Spring 2014). Retrieved from https://plato.stanford.edu/entries/concepts/#ConConAna

Martin, R., Hewstone, M., and Martin, P. Y. (2007) 'Systematic and Heuristic Processing of Majority and Minority-Endorsed Messages: The Effects of Varying Outcome Relevance and Levels of Orientation on Attitude and Message Processing', in: *Personality and Social Psychology Bulletin* 33 (1): 43–56.

McCarthy, D. (1997) 'Rights, Explanation, and Risks', in: *Ethics* 107 (2): 205–225.

McCarthy, T. (1981). The Critical Theory of Jurgen Habermas, The MIT Press, Cambridge.

McGoldrick, P. J., Keeling, K. A. and Beatty, S. F. (2008) 'A typology of roles for avatars in online retailing', in: *Journal of Marketing Management* 24 (3–4): 433–461.

Midden, C. J. H. (2011, April 29). personal communication.

Mikhail, J. (2011/2013) Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment, Cambridge University Press, Cambridge.

Mill, J. S. (1859/1985*)* On Liberty, Penguin Books, London.

Millar, J. (2014) 'Technology as moral proxy: Autonomy and paternalism by design', in: *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*: 1–7.

Miller, G. R. (2002), 'On Being Persuaded: Some Basic Distinctions', in: Dillard, J. P. and Pfau, M. (eds.), *The Persuasion Handbook: Developments in Theory and Practice,* SAGE Publications, Inc, Thousand Oaks CA.

Mintz, J. and Aagaard, M. (2012) 'The application of persuasive technology to educational settings', in: *Educational Technology Research and Development* 60 (3): 483–499.

Mitcham, C. and Schatzberg, E. (2009) 'Defining Technology and the Engineering Sciences', in: Meijers, A.W.M. (ed), *Philosophy of Technology and Engineering Sciences,* Elsevier, Amsterdam: 27–63.

Muraven, M. and Baumeister, R. F. (2000) 'Self-regulation and depletion of limited resources: Does self-control resemble a muscle?', in: *Psychological Bulletin;Psychological Bulletin* 126 (2): 247–259.

National Society of Professional Engineers. (2007). Code of Ethics, National Society of Professional Engineers. Retrieved from https://www.nspe.org/resources/ethics/code-ethics

Narita, T. and Kitamura, Y. (2010), 'Persuasive Conversational Agent with Persuasion Tactics', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), *Persuasive Technology,* Springer, Berlin/Heidelberg: 15–26).

Nass, C. and Moon, Y. (2000) 'Machines and mindlessness: Social responses to computers', in: *Journal of Social Issues* 56: 81–103.

Nelson, R. M., Beauchamp, T., Miller, V. A., Reynolds, W., Ittenbach, R. F. and Luce, M. F. (2011) 'The Concept of Voluntary Consent', in: *The American Journal of Bioethics* 11 (8): 6–16.

Nettel, A. L. and Roque, G. (2011) 'Persuasive Argumentation Versus Manipulation', in: *Argumentation* 26 (1): 55–69.

NHTSA. (2016). DOT NHTSA ODI Document - INCLA-PE16007-7876.PDF. Retrieved 13 July 2017, from https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF

Nickel, P.J. (2011) 'Ethics in e-trust and e-trustworthiness: the case of direct computer-patient interfaces', in: *Ethics and Information Technology* 13 (4): 355–363.

----- (2009) 'Trust, Staking, and Expectations.', in: *Journal for the Theory of Social Behaviour* 39 (3): 345–362.

----- (2011) 'The Definition of Technology ' (manuscript). Eindhoven.

----- (2013) 'Artificial Speech and Its Authors', in: *Minds and Machines* 23 (4): 489–502.

Nickel, P. J., Franssen, M. and Kroes, P. (2010) 'Can We Make Sense of the Notion of Trustworthy Technology?', '*Knowledge, Technology & Policy* 23 (3–4): 429–444.

Nickel, P. J., and Spahn, A. (2012), 'Trust, discourse ethics, and persuasive technology', in: Bang, M. and Ragnemalm, E. L. (eds.), *Persuasive Technology. Design for Health and Safety. The 7th International Conference. Adjunct Proceedings*, Linköping University Electronic Press, Linköping: 37–40.

Nielsen, J. (1993) Usability Engineering, Morgan Kaufmann, Amsterdam:

Nijkamp, M. D., Hollestelle, M. L., Zeegers, M. P., van den Borne, B. and Reubsaet, A. (2008) 'To be(come) or not to be(come) an organ donor, that's the question: a meta-analysis of determinant and intervention studies', in: *Health Psychology Review* 2 (1): 20–40.

Noggle, R. (1996) 'Manipulative Actions: A Conceptual and Moral Analysis', in: *American Philosophical Quarterly* 33 (1): 43–55.

Norcross, A. (2002) 'Contractualism and Aggregation', *Social Theory and Practice* 28 (2): 303–314.

Nordmann, A. (2004) Converging technologies - Research policy and organisation - EU Bookshop. Retrieved from http://bookshop.europa.eu/en/converging-technologies-pbKINA21357/

Nozick, R. (1974/2013) Anarchy, State, and Utopia, 3, Basic Books, New York.

NSW Centre for Road Safety. (2010) 'Results of the NSW Intelligent Speed Adaptation Trial Effects on road safety attitudes, behaviours and speeding', NSW Centre for Road Safety. Retrieved from http://roadsafety.transport.nsw.gov.au/downloads/isa_trial/isa_trial_results.html

NTSB. (2017). Accident ID HWY16FH018 Mode Highway occurred on May 07, 2016 in Williston, FL United. Retrieved 12 July 2017, from https://dms.ntsb.gov/pubdms/search/hitlist.cfm?docketID=59989&CFID=1126988&CFTOKEN=b1b9a5b7e849bb32-88C471C8-01C8-077F-A22C23E2E219927F

Nyholm, S. and Smids, J. (2016) 'The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?', in: *Ethical Theory and Moral Practice* 19 (5): 1275–1289.

Nyholm, S. and Smids, J. (forthcoming) 'Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and The Ethics of Mixed Traffic', in: *Ethics and Information Technology* 1–10. https://doi.org/10.1007/s10676-018-9445-9

Odom, S. L., Thompson, J. L., Hedges, S., Boyd, B. A., Dykstra, J. R., Duda, M. A., ... Bord, A. (2015) 'Technology-Aided Interventions and Instruction for Adolescents with Autism Spectrum Disorder', in: *Journal of Autism and Developmental Disorders* 45 (12): 3805–3819.

OECD and European Conference of Ministers of Transport. (2006) Speed Management, OECD Publishing.

Oei, H. and Polak, P. H. (2002) 'Intelligent Speed Adaptation (ISA) and Road Safety', in: *Journal of the International Association of Traffic and Safety Sciences (IATSS Research)* 26 (2): 41–51.

Oinas-Kukkonen, H. (2010), 'Behavior Change Support Systems: A Research Model and Agenda', in: Ploug, T. Hasle, P. and Oinas-Kukkonen, H. (eds.), *Persuasive Technology*, Springer, Berlin/Heidelberg: 4–14.

Oinas-Kukkonen, H., and Harjumaa, M. (2008), 'A Systematic Framework for Designing and Evaluating Persuasive Systems', in: Oinas-Kukkonen, H. Hasle, P. Harjumaa, M. Segerståhl, K. and Øhrstrøm, P. (eds.), *Persuasive Technology,* Springer, Berlin/Heidelberg: 164–176.

O'Keefe, D. J. (2002) Persuasion: theory and research, Sage, Thousand Oaks Calif.

Oshana, M. (2006) Personal Autonomy in Society, Routledge, Aldershot/Burlington.

Owen, G. O. (2013) 'Britain fights EU's 'Big Brother' bid to fit every car with speed limiter' in: *Mail Online*. Retrieved from http://www.dailymail.co.uk/news/article-2408012/Britain-fights-EUs-Big-Brother-bid-fit-car-speed-limiter.html

Petty, R. E. and Briñol, P. (2008a) 'Persuasion: From Single to Multiple to Metacognitive Processes', in: *Perspectives on Psychological Science* 3 (2): 137–147.

Petty, R. E. and Briñol, P. (2008b) 'Psychological Processes Underlying Persuasion A Social Psychological Approach', in: *Diogenes* 55 (1): 52–67.

Petty, R. E., Cacioppo, J. T. and Goldman, R. (1981) 'Personal involvement as a determinant of argument-based persuasion', in: *Journal of Personality and Social Psychology* 41 (5): 847–855.

Petty, R. E., Schumann, D. W., Richman, S. A. and Strathman, A. J. (1993) 'Positive mood and persuasion: Different roles for affect under high- and low-elaboration conditions', in: *Journal of Personality and Social Psychology* 64 (1): 5–20.

Ploug, T., Hasle, P. and Oinas-Kukkonen, H. (Eds.). (2010), '*Persuasive Technology*'(Vol. 6137), Springer, Berlin/Heidelberg.

Poel, I. van de. (2009), 'Values in Engineering Design', in: Meijers, A.W.M. (ed.), *Handbook of the Philosophy of Science* (Vol. 9, pp. 973–1006), Elsevier, Amsterdam: 973-1006.

----- (2011) 'The Relation Between Forward-Looking and Backward-Looking Responsibility', in: Vincent, N. A. van de Poel, I. and van den Hoven, J. (eds.), *Moral Responsibility,* Springer Netherlands: 37–52. https://doi.org/10.1007/978-94-007-1878-4_3

Poel, I. van de, Fahlquist, J. N., Doorn, N., Zwart, S. and Royakkers, L. (2011) 'The Problem of Many Hands: Climate Change as an Example', in: *Science and Engineering Ethics* 18 (1): 49–67.

Pols, A. J. K. (2012) 'How Artefacts Influence Our Actions', in: *Ethical Theory and Moral Practice* 16 (3): 575–587.

Purves, D., Jenkins, R. and Strawser, B. J. (2015) 'Autonomous Machines, Moral Judgment, and Acting for the Right Reasons', in: *Ethical Theory and Moral Practice* 18 (4): 851–872.

Qiu, L. and Benbasat, I. (2009) 'Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems', in: *Journal of Management Information Systems* 25 (4): 145–181.

Qizilbash, M. (2012) 'Informed desire and the ambitions of libertarian paternalism', in: *Social Choice and Welfare* 38 (4),: 647–658.

Railton, P. (1985), 'Locke, Stock, and Peril: Natural Property Rights, Pollution, and Risk', in: Gibson, M. (ed.), *To Breathe Freely,* Rowman and Littlefield, Tolowa NJ: 89-123.

Rawls, J. (1971). A Theory of Justice, Belknap, Harvard.

Reitberger, W., Güldenpfennig, F. and Fitzpatrick, G. (2012), 'Persuasive Technology Considered Harmful? An Exploration of Design Concerns through the TV Companion', in: Bang. M. and Ragnemalm, E. L. (eds.), *Persuasive Technology. Design for Health and Safety,* Springer, Berlin/Heidelberg: 239–250.

Reuters. (2016) 'Germany says Tesla should not use 'Autopilot' in advertising. *Reuter',* Retrieved from http://www.reuters.com/article/us-tesla-germany-idUSKBN12G0KS

Richard E. Petty, Cacioppo, J. T., Strathman, A. J. and Priester, J. R. (2005), 'To Think or Not to Think: Exploring Two Routes to Persuasion', in: Green, M. C. and Brock, T. C. (eds.), *Persuasion: Psychological Insights and Perspectives* (2nd ed.), Sage Publications, Thousand Oaks Calif.: 81–116.

Roeser, S. (2006) 'The role of emotions in judging the moral acceptability of risks', in: *Safety Science* 44 (8): 689–700.

Rosati, C. S. (1995) 'Persons, Perspectives, and Full Information Accounts of the Good', in: *Ethics* 105 (2): 296–325.

Rosén, E., Stigson, H. and Sander, U. (2011) ‘Literature review of pedestrian fatality risk as a function of car impact speed’, in: *Accident Analysis & Prevention* 43 (1): 25–33.

Roubroeks, M. (2014) Understanding social responses to artificial agents: building blocks for persuasive technology, Eindhoven University of Technology, Eindhoven.

Ruijten, P. A. M., Midden, C. J. H. and Ham, J. (2011), ‘Unconscious Persuasion Needs Goal-striving: The Effect of Goal Activation on the Persuasive Power of Subliminal Feedback’, in: *Proceedings of the 6th International Conference on Persuasive Technology: Persuasive Technology and Design: Enhancing Sustainability and Health,* ACM, New York: 4:1–4:6.

SAE International (2014) automated_driving.pdf. Retrieved 13 July 2017, from https://www.sae.org/misc/pdfs/automated_driving.pdf

SafetyNet. (2009) Cost-benefit analysis. Retrieved from https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/specialist/knowledge/pdf/cost_benefit_analysis.pdf

Saghai, Y. (2013a) ‘Salvaging the concept of nudge’, in: *Journal of Medical Ethics* 39 (8): 487-93.

Saghai, Y. (2013b). The concept of nudge and its moral significance: a reply to Ashcroft, Bovens, Dworkin, Welch and Wertheimer. *Journal of Medical Ethics* 39 (8): 499–501.

Samuels, R., Stich, S. and Bishop, M. (2002) Ending the Rationality Wars: How To Make Disputes About Human Rationality Disappear. In: Elio, R. (ed.), *Common Sense, Reasoning, and Rationality,* Oxford University Press, Oxford. DOI: http://dx.doi.org/10.1093/0195147669.003.0011

Scanlon, T. M. (2000) What we owe to each other, Belknap Press of Harvard University Press, Cambridge Mass.

----- (1982), ‘Contractualism and utilitarianism’, in: Sen, A. and Williams, B. (eds.), *Utilitarianism and Beyond,* Cambridge University Press, Cambridge/New York: 103–182.

----- (2111), ‘How I am not a Kantian’, in: Parfit, D., Scheffler, S. (ed.), *On What Matters* (Vol. 2). Oxford University Press, Oxford/New York.

Schermer, M. (2007) Gedraag je! Ethische aspecten van gedragsbeïnvloeding door nieuwe technologie in de gezondheidszorg [Behave well! Ethical aspects of influencing behavior by means of new technology in health care], Nederlandse Vereniging voor Bioethiek, Rotterdam.

Selinger, E. and Whyte, K. (2011) ‘Is There a Right Way to Nudge? The Practice and Ethics of Choice Architecture’, in: *Sociology Compass* 5 (10): 923–935.

Sher, S. (2011) ‘A Framework for Assessing Immorally Manipulative Marketing Tactics”, in: *Journal of Business Ethics* 102 (1): 97–118.

Smids, J. (2012), 'The voluntariness of persuasive technology', in: *Proceedings of the 7th international conference on Persuasive Technology: design for health and safety,* Springer-Verlag, Berlin/Heidelberg: 123–132.

Smids, J. (forthcoming) 'The Moral Case for Intelligent Speed Adaptation', in: *Journal of Applied Philosophy,* https://doi.org/10.1111/japp.12168

Sobel, D. (1994) 'Full Information Accounts of Well-Being', in: *Ethics* 104 (4): 784–810.

Somsen, H. (2011) 'When Regulators Mean Business: Regulation in the Shadow of Environmental Armageddon', in: *Netherlands Journal of Legal Philosophy* 40 (1):47-57.

Spagnolli, A., Chittaro, L. and Gamberini, L. (eds.). (2014). *Persuasive Technology* (Vol. 8462), Springer International Publishing, Cham.

Spahn, A. (2011) 'And Lead Us (Not) into Persuasion…? Persuasive Technology and the Ethics of Communication', in: *Science and Engineering Ethics* 18 (4): 633-650.

Spahn, A. (2015), 'Can Technology Make Us Happy?', in: Søraker, J. H. van der Rijt, J.W. de Boer, J.. Wong, P.H. and Brey, P. (eds.), *Well-Being in Contemporary Society,* Springer International Publishing: 93–113.

Stanovich, K. (2010) Rationality and the Reflective Mind, Oxford University Press, New York.

Steigleder, K. (2012). *Risk And Rights: Towards A Rights-Based Risk Ethics* (manuscript), Retrieved from http://www5.rz.ruhr-uni-bochum.de:8629/philosophy/mam/ethik/content/steigleder_risk_and_rights.pdf

Strack, F. and Deutsch, R. (2004) 'Reflective and Impulsive Determinants of Social Behavior', in: *Personality and Social Psychology Review* 8 (3): 220–247.

Sunstein, C. R. and Thaler, R. H. (2003) 'Libertarian Paternalism Is Not an Oxymoron', in: *The University of Chicago Law Review* 70 (4): 1159-1202.

Tajfel, H., Billig, M. G., Bundy, R. P. and Flament, C. (1971) 'Social categorization and intergroup behaviour', in: *European Journal of Social Psychology* 1 (2): 149–178.

Teuber, A. (1990) 'Justifying Risk', in: *Daedalus'* 119 (4): 235–254.

Thaler, R. H. and Sunstein, C. R. (2009) Nudge: Improving Decisions About Health, Wealth, and Happiness (Revised and Expanded). Penguin Books.

The Tesla Team. (2016) 'A Tragic Loss | Tesla', Retrieved 12 July 2017, from https://www.tesla.com/blog/tragic-loss.

Thomson, J. J. (1985a), 'Imposing risks', in: Gibson, M. (ed.), *To Breathe Freely.* Rowman and Allanheld, Totowa N.J.: 124–140.

----- (1985b) 'The Trolley Problem', in: *The Yale Law Journal* 94 (5): 1395–1515.

----- (2008) 'Turning the Trolley', in: *Philosophy and Public Affairs* 36 (4): 359–374.

to persuade. (1988). *Webster's New World Dictionary, 3rd College Edition*. Simon & Schuster.

to persuade. (2011). *Merriam Webster Online Dictionary*.

Todd, P. M. and Gigerenzer, G. (2000) 'Précis of Simple heuristics that make us smart', in: *Behavioral and Brain Sciences* 23 (05): 727–741.

Todorov, A., Chaiken, S. and Henderson, M. D. (2002), 'The heuristic-systematic model of social information processing', in: Dillard, J. P. and Pfau, M. (eds.), *The Persuasion Handbook: Developments in Theory and Practice*, SAGE Publications, Inc, Thousand Oaks CA: 195–211.

Torning, K. and Oinas-Kukkonen, H. (2009), 'Persuasive system design: state of the art and future directions', in: *Proceedings of the 4th International Conference on Persuasive Technology*, ACM, New York: 30:1–30:8).

Tromp, N., Hekkert, P. and Verbeek, P.P. (2011) 'Design for Socially Responsible Behavior: A Classification of Influence Based on Intended User Experience', in: *Design Issues* 27 (3): 3–19.

Trout, J. D. (2005) 'Paternalism and Cognitive Bias', in: *Law and Philosophy* 24 (4): 393–434.

Tversky, A. and Kahneman, D. (1974) 'Judgment under Uncertainty: Heuristics and Biases', in: *Science* 185 (4157): 1124–1131.

van Baaren, R. B., Holland, R. W., Steenaert, B. and van Knippenberg, A. (2003) 'Mimicry for money: Behavioral consequences of imitation', in: *Journal of Experimental Social Psychology* 39 (4): 393–398.

van der Pas, J. W. G. M., Marchau, V. A. W. J., Walker, W. E., van Wee, G. P., and Vlassenroot, S. H. (2012) 'ISA implementation and uncertainty: A literature review and expert elicitation study', in: *Accident Analysis & Prevention* 48: 83–96.

van Loon, R. J. and Martens, M. H. (2015) 'Automated Driving and its Effect on the Safety Ecosystem: How do Compatibility Issues Affect the Transition Period?', in: *Procedia Manufacturing* 3: 3280–3285.

Vanderheiden, S. (2006) 'Assessing the case against the SUV', in: *Environmental Politics* 15 (1): 23–40.

Verbeek, P.-P. (2006). *Persuasive Technology and Moral Responsibility. Toward an ethical framework for persuasive technology*. Retrieved from https://www.utwente.nl/bms/wijsb/organization/verbeek/verbeek_persuasive06.pdf

----- (2009) 'Ambient Intelligence and Persuasive Technology: The Blurring Boundaries Between Human and Technology', in: *NanoEthics* 3 (3): 231–242.

Verberne, F. M. F. (2015) Trusting a virtual driver. Similarity as a trust cue, Eindhoven University of Technology, Eindhoven.

Verberne, F. M. F., Ham, J., Ponnada, A. and Midden, C. J. H. (2013) ' Trusting Digital Chameleons: The Effect of Mimicry by a Virtual Social Agent on User Trust', in: *Persuasive Technology,* Springer, Berlin/Heidelberg: 234–245.

Verwey, W. B., Brookhuis, K. A. and Janssen, W. H. (1996). Safety effects of in-vehicle information system (No. TM-96-C00). TNO Human Factors, Soesterberg.

Vlassenroot, S., Broekx, S., Mol, J. D., Panis, L. I., Brijs, T. and Wets, G. (2007) 'Driving with intelligent speed adaptation: Final results of the Belgian ISA-trial', in: *Transportation Research Part A: Policy and Practice* 41 (3): 267–279.

Vlassenroot, S., Marchau, V., De Mol, J., Brookhuis, K. and Witlox, F. (2011) 'Potential for in-car speed assistance systems: results of a large-scale survey in Belgium and the Netherlands', in: *IET Intelligent Transport Systems* 5 (1): 80–89.

Voerman, S. (2015), 'Tackling your lifestyle with the body coach', in: Kool, L. Timmer, J. and van Est, R (eds.), *Sincere Support. The Rise of the E-coach.* Rathenau Instituut, Den Haag: 34-69.

Wallach, W. and Allen, C. (2009) Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, Oxford.

Wang, L. C., Baker, J., Wagner, J. A. and Wakefield, K. (2007) 'Can a Retail Web Site Be Social?', in: *Journal of Marketing* 71 (3): 143–157.

Warner, T. (2012) 'E-coaching systems: Convenient, anytime, anywhere, and nonhuman', in: *Performance Improvement* 51 (9): 22–28.

Wilkinson, T. M. (2013) 'Nudging and Manipulation', in: *Political Studies* 61 (2): 341–355.

Windsor, M. (2015) 'Will your self-driving car be programmed to kill you if it means saving more strangers?' Retrieved 19 February 2016, from https://www.sciencedaily.com/releases/2015/06/150615124719.htm

Wolf, G. (2010) 'The Data-Driven Life', in: *The New York Times.* Retrieved from http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html

Wood, A. (2011) 'Humanity as an End in Itself', in: Parfit, D., Scheffler, S. (ed.), *On What Matters* (Vol. 2). Oxford University Press.

Worstall, T. (2014) 'When Should Your Driverless Car From Google Be Allowed To Kill You?', in: *Forbes.* Retrieved from http://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you/

Yadron, D. and Tynan, D. (2016) 'Tesla driver dies in first fatal crash while using autopilot mode', in: *The Guardian.* Retrieved from https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk

Yeung, K. (2011) 'Can We Employ Design-Based Regulation While Avoiding Brave New World?', in: *Law, Innovation and Technology* 3 (1): 1–29.

Zajonc, R. B. (2001) 'Mere Exposure: A Gateway to the Subliminal', in: *Current Directions in Psychological Science* 10 (6): 224–228.

Zimmerman, D. (2010) 'Taking Liberties', in: *Social Theory and Practice* 28 (4): 577–609.

Summary

Persuasive Technology, Allocation of Control, and Mobility. An Ethical Analysis

Persuasive technology is technology that is intentionally designed to change the mental states and the behavior of its users. As is argued in this thesis, it is characteristic of this technology that, although it 'nudges' users in a certain direction, it leaves them substantial control over what they think and do. The main part of this thesis is devoted to an investigation of some central ethical aspects of persuasive technology. Mobility is the main application context, but to a lesser extent, cases come from health care, commerce, and other contexts. A smaller part of the thesis investigates ethical aspects of mobility-related technologies that take over full control from its users, such as autonomous cars.

In this thesis, several methodologies are combined, of which conceptual analysis is the main one. Much attention is given to 'allocation of control', because it is ethically crucial how control is distributed between users on the one hand, and persuasive technology and automation technology on the other hand. At the same time, the thesis is empirically informed. The research was carried out in the context of an NWO-funded interdisciplinary research project that involving a psychology PhD project, in which experiments were carried out with human subjects. The results of this latter project are reflected upon in Chapter 5. The thesis makes extensive use of the psychological literature in order to understand the underlying psychological mechanisms of the persuasive technology's influence methods. Many references are made to (prototypes of) concrete examples of persuasive technology, and the thesis contains several smaller and two large case studies (on Intelligent Speed Adaptation, and on 'similarity-based influence').

Chapter 2 summarizes and extends the existing literature on the ethics of persuasive technology. Its main contributions are twofold. First, it disambiguates the concept of "outcome" into two elements: the behavior that results from user interaction with the persuasive technology on the one hand, and how that behavior has an impact on central values, e.g. health, sustainability, and the like, on the other hand. Second, instead of looking at designers alone, it looks at the interaction between designers and deployers of persuasive technology, which is shown to confer additional insights.

Chapter 3 provides a broader background to the ethics of persuasive technology by contrasting a related and partly overlapping concept, 'nudging', with rational persuasion. The psychological processes underlying persuasion are studied in order to get grips on the extent to which 'non-argumentative means of persuasion' grant the user of persuasive technology substantial control over her mental states and behavior. Three guidelines are proposed for guaranteeing such control, and applied to contrast nudging with persuasion.

Chapter 4 builds on the results of chapter 3 in order to redefine persuasive technology: it is technology that influences by means of communicating with users in a way that grants them substantial control. This improved definition allows one to better distinguish persuasive technology from manipulation and coercion.

Whereas Chapters 2, 3 and 4 are of a more general and theoretical nature, Chapters 5-8 are rather of an applied nature.

Chapter 5 zooms in on the use of one particular type of 'non-argumentative means of persuasion', viz., similarity-based influence (which was studied by the psychology PhD-student in the NWO project). This concerns ways in which the persuasive technology, often operating under the guise of artificial social agents, is made to appear similar to its users, such as mimicking the user's speech pattern and head-movement. Three guidelines are developed for responsible use of similarity-based influence.

Chapter 6 deals with the ethics of risk by way of preparation for Chapters 7 and 8. The chapter discusses how Scanlon's contractualism deals with the question of acceptable risk-imposition. It argues that Scanlon's commitment to avoid inter-personal aggregation leads to very stringent moral restrictions, on which risk impositions are morally acceptable. The discussion of this chapter provides some important normative background for Chapters 7 and 8, in which the ethics of risk is a key issue.

Chapter 7 argues for a positive moral case for Intelligent Speed Adaption, based on its potential to significantly reduce driving risks and the resulting harms. In the course of doing so, it contrasts a persuasive version of Intelligent Speed Adaption (which warns when speeding) with a limiting form (which makes speeding technologically impossible). It concludes that objections to both are unconvincing and that governments would do well to start with mandating the use of the persuasive variant.

Chapter 8 argues against conceiving of the problem of programming autonomous cars for situations of unavoidable accident as an applied trolley problem. It does so on the basis of three major disanalogies. First, the decision-making situations are very different. Second, unlike the trolley problem, programming autonomous cars involves various considerations of existing moral and legal responsibilities. Third, whereas the description of trolley scenario's consists of stipulated facts and certainties, any plausible scenario of an accident of autonomous cars is characterized by risks, probabilities, and uncertainties.

Chapter 9 summarizes the main conclusions of the thesis and points out several avenues for further study.

Samenvatting

Persuasieve technologie, allocatie van controle en mobiliteit. Een ethische analyse.

Persuasieve technologie is technologie die doelbewust ontworpen is om wat gebruikers denken en doen te veranderen. Zoals in dit proefschrift wordt betoogd, is het karakteristiek voor persuasieve technologie dat, hoewel zij haar gebruikers in een bepaalde richting 'duwt', gebruikers toch substantiële controle houden over hun mentale toestanden en hun gedrag, over wat ze denken en doen. Het grootste deel van dit proefschrift is gewijd aan een onderzoek naar enkele cruciale ethische aspecten van persuasieve technologie. Hierbij is mobiliteit de belangrijkste gebruikscontext, maar ook voorbeelden uit de gezondheidszorg, commercie en andere contexten komen aan bod. Het laatste deel van het proefschrift onderzoekt enkele ethische aspecten van robottechnologie in het domein van mobiliteit, namelijk zelfrijdende auto's.

In dit proefschrift combineer ik verschillende onderzoeksmethoden, waarvan begripsanalyse de belangrijkste is. Veel aandacht is er voor 'allocatie van controle', omdat het in ethisch opzicht cruciaal is hoe de controle verdeeld wordt tussen gebruikers enerzijds en persuasieve- en robottechnologie anderzijds. Deze toegepaste begripsanalyse is tegelijkertijd geïnformeerd door wetenschappelijke kennis en de praktijk. Het onderzoek werd namelijk uitgevoerd als onderdeel van een door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek gefinancierd interdisciplinair onderzoeksproject. Hierin waren 'automotive'- en 'mens-techniek-interactie' onderzoekers betrokken en deed een promovendus van psychologie experimenteel onderzoek waarin proefpersonen persuasieve technologie gebruikten. Hoofdstuk 5 bevat een ethische reflectie op de uitkomsten van dit onderzoek. Verder maak ik in het proefschrift uitgebreid gebruik van de psychologische vakliteratuur om de onderliggende mechanismen van de methoden waarmee persuasieve technologie probeert te beïnvloeden te begrijpen. Tot slot komen er vaak concrete persuasieve technologieën (of prototypen daarvan) aan bod en ook bevat het proefschrift verscheidene 'case studies' (o.a. over 'invloed gebaseerd op gelijkenis' en over Intelligente Snelheids Adaptatie).

Na de inleiding in hoofdstuk 1 geeft hoofdstuk 2 een samenvattend overzicht van de bestaande literatuur over de ethiek van persuasieve technologie. Voortbouwend hierop beargumenteer ik de waarde van twee uitbreidingen. In de eerste plaats is het verhelderend om twee betekenissen van het begrip ‘uitkomst’ te onderscheiden: enerzijds het *gedrag* dat het resultaat is van de interactie tussen persuasieve technologie en gebruiker en anderzijds de *gevolgen van dat gedrag* voor belangrijke waarden, zoals gezondheid en duurzaamheid. In de tweede plaats kunnen we in plaats van alleen naar ontwerpers beter kijken naar het samenspel tussen de ontwerpers en de partij die de persuasieve technologie wil gaan gebruiken om bepaald gedrag van mensen te veranderen. Dit levert namelijk nieuwe inzichten op, onder andere met betrekking tot de vraag naar de verantwoordelijkheden van verschillende partijen betrokken bij persuasieve technologie.

Hoofdstuk 3 geeft een bredere achtergrond aan de ethiek van persuasieve technologie door een gerelateerd en gedeeltelijk overlappend concept, namelijk ‘nudging’ (letterlijk een ‘duwtje’ en de term voor een subtiele vorm van invloed die recent veel aandacht krijgt), te contrasteren met rationele persuasie (wat ongeveer ‘overtuigen’ betekent). In rationale persuasie spelen argumenten de hoofdrol, maar kunnen ook niet-argumentatieve overtuigingsmiddelen zoals emoties, of ‘een betrouwbare uitstraling’ meedoen. De psychologische processen die aan persuasie ten grondslag liggen worden bestudeerd om te beoordelen in hoeverre deze niet-argumentatieve overtuigingsmiddelen de persoon die onderwerp is van rationele persuasie substantiële controle over zijn mentale toestanden en handelen doet behouden. Om die controle te garanderen formuleer ik drie richtlijnen voor het gebruik van niet-argumentatieve overtuigingsmiddelen en op basis van deze richtlijnen vergelijk ik ‘nudging’ en rationele persuasie.

Voortbouwend op hoofdstuk 3 wordt in hoofdstuk 4 een herdefinitie van persuasieve technologie gegeven: het is technologie die beïnvloedt door te communiceren met gebruikers op een wijze die hen substantiële controle laat over wat ze denken en doen. Dat betekent dat de gebruiker voldoende ruimte krijgt om, desgewenst, zelf na te denken over zijn redenen voor verschillende handelingsmogelijkheden en daar vervolgens ook naar te handelen. Ik werk deze substantiële controle uit in drie voorwaarden. Ten eerste dient de persuasieve technologie op zo’n wijze met de gebruiker te communiceren dat die voldoende informatie krijgt, bijvoorbeeld over wat het doelgedrag precies is, of bruikbare feedback op het eigen handelen. Ten tweede mag de persuasieve technologie niet tegelijkertijd dit praktisch redeneren van de gebruiker omzeilen of verstoren, bijvoorbeeld door misleidende feedback of groepsdruk. Ten derde dient de gebruiker vrij te zijn in zijn uiteindelijke handelen. De persuasieve technologie mag het bijvoorbeeld de gebruiker niet heel moeilijk maken om anders te handelen dan de technologie 'wil'. Gebruik makend van de verbeterde definitie kunnen we scherper onderscheiden tussen persuasieve technologie en technologie die veeleer manipuleert, dwingt, of bepaald gedrag simpelweg technologisch onmogelijk maakt. Zo kwalificeren gordelverklikkers zich niet als persuasieve technologie, omdat de pieptonen door de meeste gebruikers als zeer irritant en daarmee als dwingend ervaren worden; aan de derde voorwaarde wordt dus niet voldaan.

Waar de hoofdstukken 2, 3 en 4 meer van een algemene en theoretische aard zijn, hebben de hoofdstukken 5 t/m 8 een meer toegepast karkater. In hoofdstuk 5 wordt een gedetailleerde ethische analyse gegeven van een specifieke beïnvloedingstrategie, namelijk invloed gebaseerd op gelijkenis. Dit betreft de manier waarop ontwerpers persuasieve technologie, vaak in de vorm van digitale sociale 'agents', laten lijken op gebruikers. Zo kan de digitale 'agent' bijvoorbeeld zijn gezicht laten lijken op dat van een gebruiker, diens hoofdbewegingen nadoen, of diens spreekpatroon imiteren. Het gevaar hierbij is groot dat de gebruiker gemanipuleerd wordt: de persuasieve sociale 'agent' beïnvloedt de gebruiker op een manier die moeilijk op te merken en te controleren is. Drie richtlijnen worden ontwikkeld voor een ethisch verantwoord gebruik van invloed gebaseerd op gelijkenis.

Hoofdstuk 6 handelt over risico-ethiek, ter voorbereiding op hoofdstuk 7 en 8. Besproken wordt hoe het contractualisme van Thomas Scanlon om gaat met de vraag wanneer en waarom het acceptabel is om risico's voor anderen dan jezelf te veroorzaken. Ik betoog dat Scanlon's commitment om in zijn ethische theorie geen beroep te doen op interpersoonlijk aggregatie (ongeveer het optellen van de baten voor vele individuen, ten gevolge van bijvoorbeeld een beleidsbeslissing om een bepaald risico toe te staan, om die vervolgens af te wegen tegen de opgetelde lasten voor anderen) leidt tot zeer strikte beperkingen aan welke risico's acceptabel zijn. De discussie in dit hoofdstuk bereidt voor op hoofdstuk 7 en 8 waarin risico-ethiek een sleutelrol speelt.

In hoofdstuk 7 betoog ik eerst dat de huidige verkeersrisico's moreel niet acceptabel zijn, met name niet voor kwetsbare verkeersdeelnemers zoals voetgangers en fietsers. Dit omdat de risico's hoger zijn dan nodig (met name door te hard rijden), omdat de risico's onrechtvaardig verdeeld zijn en ook omdat er meer haalbare voorzorgsmaatregelen genomen zouden kunnen worden dan nu het geval is. De grote veiligheidswinst die we in het verkeer kunnen boeken met Intelligente Snelheids Adapatatie (ISA) is dan ook voldoende reden om deze technologie verplicht te stellen voor alle auto's. Dit geld zowel voor persuasieve versies van ISA, die vooral waarschuwen bij te hard rijden, als limiterende versies, die te hard rijden technisch onmogelijk maken. Verschillende bezwaren tegen deze technologie blijken bij nadere inspectie niet overtuigend. Deze voorbeeldstudie naar ISA illustreert hoe preventie van ernstige schade aan derden voldoende reden kan zijn om het gebruik van persuasieve technologie verplicht te stellen.

In hoofdstuk 8 betogen Sven Nyholm en ik dat het probleem van hoe zelfrijdende auto's geprogrammeerd dienen te worden voor situaties waarin een ongeluk onvermijdelijk is beter niet benaderd kan worden als een toegepast 'trolley probleem'. Het 'trolley probleem' is een beroemd filosofisch gedachtenexperiment dat draait om de vraag of en wanneer we het geoorloofd vinden om het leven van één persoon op te offeren om dat van vijf anderen te redden. We laten zien dat er drie belangrijke verschillen zijn tussen het programmeren van 'ongeluk-algoritmes' en het 'trolley probleem'. In de eerste plaats zijn de situaties waarin de beslissing genomen moet worden heel verschillend. In het trolley dilemma moet één persoon in een ogenblik beslissen, terwijl er verschillende partijen betrokken zijn bij het bepalen hoe zelfrijdende auto's geprogrammeerd dienen te worden, die daar ruim de tijd voor kunnen nemen. In de tweede plaats nemen die partijen, anders dan in het trolley probleem, verschillende van toepassing zijnde morele en wettelijke verantwoordelijkheden mee in hun overwegingen. Ten derde zijn in het trolley probleem alle omstandigheden duidelijk en de gevolgen van de verschillende handelingsmogelijkheden vooraf met zekerheid bekend. Daarentegen wordt elk realistisch scenario van een botsing met een zelfrijdende auto gekenmerkt door risico's, waarschijnlijkheden en onzekerheden. Een geloofwaardige risico-ethiek is daarom onmisbaar voor het programmeren van zelfrijdende auto's.

In hoofdstuk 9 vat ik de belangrijkste conclusies van het proefschrift samen, geef ik een aantal vragen voor vervolgonderzoek en doe ik een aantal aanbevelingen voor beleid rondom persuasieve technologie.

Curriculum Vitae

Jilles Smids was born on May 23rd, 1976 in Apeldoorn, the Netherlands. He studied chemistry at Utrecht University (MSc, 2002) and philosophy at the same university (MA, 2011). Before starting his PhD project, he worked as a chemistry and a philosophy teacher at schools for secondary education From 2010 to 2015 he carried out his PhD research at Eindhoven University of Technology, as part of the interdisciplinary NWO-funded project “Persuasive Technology, Allocation of Control, and Social Values”. His work has a focus on the ethics of persuasive technology, the ethics of risk, and the ethics of mobility.

From the academic year 2015/16 onwards, Jilles has been a chemistry teacher at Calvijn College in Goes, the Netherlands. He is married to Marian Vader and they have three children: Simon, Maria, and Job.

Simon Stevin Series in Ethics of Technology
Delft University of Technology, Eindhoven University of Technology,
University of Twente & Wageningen University
Editors: Philip Brey, Anthonie Meijers and Sabine Roeser

Books and Dissertations

Volume 1: Lotte Asveld, '*Respect for Autonomy and Technology Risks*', 2008

Volume 2: Mechteld-Hanna Derksen, *'Engineering Flesh, Towards Professional Responsibility for 'Lived Bodies' in Tissue Engineering'*, 2008

Volume 3: Govert Valkenburg, *'Politics by All Means. An Enquiry into Technological Liberalism'*, 2009

Volume 4: Noëmi Manders-Huits, *'Designing for Moral Identity in Information Technology'*, 2010

Volume 5: Behnam Taebi, *'Nuclear Power and Justice between Generations. A Moral Analysis of Fuel Cycles'*, 2010

Volume 6: Daan Schuurbiers, *'Social Responsibility in Research Practice. Engaging Applied Scientists with the Socio-Ethical Context of their Work'*, 2010

Volume 7: Neelke Doorn, *'Moral Responsibility in R&D Networks. A Procedural Approach to Distributing Responsibilities'*, 2011

Volume 8: Ilse Oosterlaken, *'Taking a Capability Approach to Technology and Its Design. A Philosophical Exploration'*, 2013

Volume 9: Christine van Burken, *'Moral Decision Making in Network Enabled Operations'*, 2014

Volume 10: Faridun F. Sattarov, *'Technology and Power in a Globalising World, A Political Philosophical Analysis'*, 2015

Volume 11: Gwendolyn Bax, *'Safety in large-scale Socio-technological systems. Insights gained from a series of military system studies'*, 2016

Volume 12: Zoë Houda Robaey, *'Seeding Moral Responsibility in Ownership. How to Deal with Uncertain Risks of GMOs'*, 2016

Volume 13: Shannon Lydia Spruit, *'Managing the uncertain risks of nanoparticles. Aligning responsibility and relationships'*, 2017

Volume 14: Jan Peter Bergen, '*Reflections on the Reversibility of Nuclear Energy Technologies*', 2017

Volume 15: Jilles Smids, *'Persuasive Technology, Allocation of Control, and Mobility. An Ethical Analysis'*, 2018

Simon Stevin (1548-1620)

'Wonder en is gheen Wonder'

This series in the philosophy and ethics of technology is named after the Dutch / Flemish natural philosopher, scientist and engineer Simon Stevin. He was an extraordinary versatile person. He published, among other things, on arithmetic, accounting, geometry, mechanics, hydrostatics, astronomy, theory of measurement, civil engineering, the theory of music, and civil citizenship. He wrote the very first treatise on logic in Dutch, which he considered to be a superior language for scientific purposes. The relation between theory and practice is a main topic in his work. In addition to his theoretical publications, he held a large number of patents, and was actively involved as an engineer in the building of windmills, harbours, and fortifications for the Dutch prince Maurits. He is famous for having constructed large sailing carriages.

Little is known about his personal life. He was probably born in 1548 in Bruges (Flanders) and went to Leiden in 1581, where he took up his studies at the university two years later. His work was published between 1581 and 1617. He was an early defender of the Copernican worldview, which did not make him popular in religious circles. He died in 1620, but the exact date and the place of his burial are unknown. Philosophically he was a pragmatic rationalist for whom every phenomenon, however mysterious, ultimately had a scientific explanation. Hence his dictum 'Wonder is no Wonder', which he used on the cover of several of his own books.