Toward a Philosophy and Ethics of Persuasive Technology

Jilles Smids

Master's Thesis, Utrecht University / Eindhoven University of Technology

August 2011

Foreword

This master's thesis was written while I was a beginning PhD-student, working on a four-year project on the ethics of persuasive technology, at Einhoven University of Technology. I have experienced it as a wonderful opportunity to write my master's thesis in such a stimulating research community as I would say our Eindhoven *Philosophy and Ethics* section is. It implied a process consisting of cycles of writing, criticism, and rewriting, that I often longed for when I was only a student. At the same time, I came to realize that academic writing is much harder than writing papers for philosophy courses, which at most included feedback on the set-up of the paper. I can only hope that all these efforts have resulted in a convincing thesis.

Meanwhile, writing this thesis has learned me how interesting the phenomenon of persuasive technology is. I am happy to know that I have more than three years to work on issues left undecided in this thesis, and on all the other interesting avenues for further research that became visible.

I would like to thank the people that have helped me during my writing of this thesis. First of all, I thank my supervisor on behalf of the Eindhoven University of Technology (both for this thesis and for my PhD-project) Andreas Spahn for many valuable discussions and continuing enthusiastic support. Jan Vorstenbosch is thanked for being supervisor on behalf of the Department of Philosophy of Utrecht University. He has made many valuable comments on this thesis and suggestions for future research. Frank Verberne, my fellow PhD in the project, researching the psychology of persuasion, helped me considerably by our stimulating and insight giving discussions. Thanks to Philip Nickel and Anthonie Meijers for valuable comments on chapter two. Jaap Ham gave helpful literature references, and reflections on the field of study of persuasive technology. Cees Midden has made several insight giving comments during the project-meetings, which often made me thinking for a longer period. Again, it is a privilege to conduct my PhD research in such an environment, and I am looking forward to the coming years.

Table of contents

1 Introduction

1.1 What is persuasive technology?

Persuasion is an age-old phenomenon and it comes therefore as no surprise that in an increasingly technological life-world, persuasive methods were built into technology, giving rise to the phenomenon of persuasive *technology*. Persuasive technologies are intentionally designed to change attitude, behavior, or both [Fogg, 2003, Ch. 1]. For example, an electronic display next to a road that displays the text message "you are driving too fast", accompanied by a thrown (sad face). Or, software designed to change user behavior in such ways that users of computers do not get RSI. Commercial websites like amazon.com provide the user with tailored suggestions about books she might want to buy as well (based on actual search, and purchasing history), or recommend books on the basis of experts, or peers. In order to reduce teenage pregnancy, the simulator *Baby Think It Over*[1] was developed. This is a high-tech realistic looking baby doll that has an embedded computer that creates crying sounds at random times. The teenager has to give appropriate care to the doll in order to stop it crying [Fogg, 2003, pp. 78-79]. Several websites exist which aim to support people to change their eating behavior and to reduce weight. Such websites typically involve feedback on calorie intake, provide suggestions for food and exercising, social support from peers, and e-coaching by experts [Letho & Oinas-Kukkonen 2010]. More and more cars are equipped with eco-dashboards, that assist and persuade drivers to change their driver behavior in ways that enhance fuel-economy; an example is the Honda Ecological Drive Assist System[2].

As a distinct research discipline, persuasive technology (hereafter PT) originated in the 1990's, with important roots in the field of human-computer interaction. B.J. Fogg is the most important founder of the discipline and his book on persuasive technology [2003] is up to now the only monograph on PT. In 2006, the first annual scientific conference on PT was held [Ijselstein et. al.]. What becomes salient from Fogg's book, from reflections given by prominent PT-scholars, and from personal reports on the annual PT conference[3], is the general acknowledgement in the field that ethical reflection on PT is highly desired. To this day, only a handful initial attempts of an 'ethics of PT' are available and, as I will argue in this thesis, these attempts need revision and extension.

[1] A similar version is the RealCareBaby, see http://www.realityworks.com/infantsimulations/realcarebaby.asp, accessed at June 28, 2011

[2] http://world.honda.com/news/2008/4081120Ecological-Drive-Assist-System/ accessed on 17 august 2011

[3] Personal communication with Jaap Ham, and Andreas Spahn on several occasions during 2011

Why is ethical reflection so needed? The aim of PT is to influence the behavior of users into directions desired by the designers and the employers of the PT. This seems directly to conflict with the most cherished value of Western civilization: autonomy. We generally don't like paternalism, and PT seems be a technological source of paternalism. However, if PT relies on voluntary change, as leading PT scholars emphasize (see chapter two), then how is autonomy affected? One important focus for ethical reflection on PT therefore is the development of methods to assess the *actual* voluntariness of behavior change. Other ethical concerns have to do with designer responsibilities, with the ethical acceptability of the aims of persuasion, and with the political question of when to enforce the use of PT in certain contexts. *This thesis aims to contribute to the development of a convincing framework for ethical reflection on PT, and, foundationally to such a framework, it also aims to present a fist attempt of a philosophy of PT, i.e. a conceptual analysis of the phenomenon of PT*. In the next section, I will provide an outline.

1.2 Outline of this thesis.

As said above, the two core questions of this thesis are: what is PT and how can we proceed to ethically design and evaluate PT? I need first give attention to the former question, in order to be able to say anything relevant about the latter question. Without a conceptual analysis of the concept of PT, it is unclear what the domain of ethical reflection is, and furthermore, it is not possible to distinguish (technological) persuasion from other forms of influencing (with or without the usage of technology), such as convincing, incentivizing, manipulating, and coercing, which all have different ethical implications.

Whereas the ethical reflection on PT is still in its infancy, philosophical reflection on PT is nearly absent. Therefore, in chapter two, I will first try to give a conceptual analysis, mainly by searching for a core or an underlying principle of persuasion that applies also to PT. This principle is formulated from the perspective of a prominent approach to the psychology of persuasion, viz. the elaboration likelihood model (ELM, section 2.2): any attempt of persuasion should allow for the *possibility* of recipients to assess their reasons for compliance with the target change by engaging in high amounts of issue-relevant thinking (section 2.3). This underlying principle can account for the broadness of the phenomenon of persuasion, and furthermore, it enables to make the distinction with the other forms of influencing mentioned above. With this principle as its core, a definition of PT is provided and tested on a number of cases (section 2.4). In section 2.5, I discuss some further issues with regard to the definition.

In chapter 3, which is on the ethics of PT, I will start with an examination of four existing initial attempts to develop a framework for ethical reflection on PT (section 3.1). By clarifying

and extending elements of these frameworks, I propose a new framework, of which the core is formed by i) methods of technological persuasion, ii) target change (i.e. behavior or its mental determinants), and iii) the final end served by that target change (section 3.2) For example, a weight-loss website (methods of persuasion) persuades users to exercise more regularly (target change), which leads to increased health (final end). By discussing several elements of the framework in their mutual relation, I hope to shed light on what ethical issues relate to PT and on how they could be answered. In this discussion, the emphasis will lie on persuasive methods, because the conceptual clarification of PT in chapter two is mainly about the methods of technological persuasion (for these methods make a technological artifact an instance of PT). It turns out that psychological knowledge of persuasion, i.e. the ELM of persuasion, rather naturally generates ethical questions about PT.[4] Whereas section 3.2 develops a framework for ethical reflection, and is thus about the right procedure, section 3.3 , finally, will provide somewhat more substantial ethical guidance; it is devoted to a treatment of the relation between PT and autonomy.

[4] I take this as an important virtue of the ELM, because good theories help to ask relevant and interesting research questions.

2 Conceptual clarification of persuasive technology

2.1 Introduction: the standard definition and its shortcomings

Persuasive technology as a distinct research discipline is rather young, approximately something more than a decade now. As a consequence, definitional and conceptual issues have not received much attention until now. Considered the complexities involved in defining persuasive technology and the problems of the currently most prominent definition, this lack of attention is regrettable. The aim of this chapter is to remedy this situation, by proposing another definition and subsequently showing that this definition does not suffer from the same problems. In this introduction, I will first diagnose the problems with the present standard definition and then outline the structure of this chapter.

If we combine what Fogg [2003, p 1, pp.15,16] says about the definition of PT, we arrive at:

> *PTs are technologies[5] which are intentionally designed to change people's behavior, attitude or both (without using coercion or deception; persuasion implies voluntary change).*

Unsurprisingly considered the status of B.J. Fogg in the field of study of PT, this definition has become more or less the standard definition. This is evident from for example Ijsselsteijn et. al. [2006, 1] and Oinas-Kukkonen [2010, p6]. 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[6]". He goes on to state that "persuasion relies on the user's voluntary participation in the persuasion process", thus agreeing on Fogg's addition, as placed between the brackets in the above definition..

The inadequacy of this definition becomes visible from the fact that its scope is too wide: it includes technologies which are intuitively no PTs, for example the handle by which you can open a door counts as a PT[7], because it is intentionally designed to change the behavior of a person (Latour [1992] would say: change relative to making a hole in the door) in a way that depends on voluntary action of that person. It seems however that a handle is

[5] For present purposes, I will pay no attention to defining 'technology', because this is a project of its own.(See for example Mitscham & Schatzberg [2009])], and in the far majority of cases of PT, we are able to see that we deal with a technological artefact without needing a definition of technology.
[6] By inducements he means economic incentives.
[7] I owe this example to Frank Verberne

not a PT in the proper sense and a better definition should make clear why not. Another example [Nickel[8], work in progress] is a belief-or-behavioral disposition inducing pill, offered and accepted voluntarily. This pill would on the standard definition also count as a PT but seems to be none.

The diagnosis of this problem with the scope is the definition's lack of the specification of an underlying principle or mechanism of persuasion [Nickel, p 1]. The definition does not specify what is characteristic of the concept of persuasion. Fogg (and the scholars following his definition) seems to be aware of the absence of a principle of persuasion, because he puts limitations on which methods count as persuasive: deception (an informational form of manipulation) and coercion are explicitly ruled out and a voluntariness condition is added. This will not do however, because i) besides manipulation and coercion, more 'mechanisms of influence' exist, for example incentivizing and convincing, and ii) a positive characterization of persuasion is still needed in order to judge its voluntariness and iii) a definition of PT should clarify the conceptual unity of its instances, for which a principle of persuasion is indispensable.

It is probably not accidental that a principle or mechanism of persuasion is absent in the standard definition. Persuasion is a very broad and perhaps diffuse phenomenon and the concept is notoriously difficult to characterize. Still, a better definition is desirable for a number of reasons [Nickel, pp 1-2]. In the first place, the study and design of PTs is facilitated by clear distinction between "persuasion and other mechanisms of changing people's attitudes and behavior". In the second place, ethical reflection on PTs involves an assessment of the actual voluntariness of persuasion, for which knowledge of the mechanism of persuasion is needed. Finally, definitional clarity may prevent different views on what persuasion is between users and designers and thus prevent a general distrust of PTs.

In the remainder of this chapter, I will proceed as follows. In section 2.2, I will investigate the psychological mechanisms underlying persuasion, as studied in persuasion research. This way to start is motivated by the broadness and complicatedness of the phenomenon of persuasion that makes a conceptual philosophical analysis of 'persuasion' apparently more difficult then 'coercion' or 'manipulation'. Thus, I seek empirical input; in particular, I employ a prominent theoretical perspective on persuasion, the so-called elaboration likelihood model (ELM). Importantly, this model views persuasion as a form of communication, which already substantially narrows down the concept of persuasion.

Using the ELM, I develop my central idea that persuasion is characterized by the fact that it *enables, or allows for*, change of mental states that determine behavior through what

[8] In writing this chapter, my thinking was considerably furthered by Nickel's manuscript, although I disagree with him on important points. I will specify places in which I explicitly take up his insights.

the ELM calls the central route to persuasion. That means that the recipient has the *possibility* to engage in high amounts of issue-relevant thinking, and in that way to find out whether they have reason to change in the way intended by the persuader. This is all not to say that the *actual* change will be proceed via the central route, because the actual mechanism or principle of change, according to the ELM, depends on the way the recipient processes the message.

As I will argue in section 2.3, the broadness of the phenomenon of persuasion originates from this diversity of what can actually happen in the mind of the recipient. However, the *possibility* of central route change is what distinguishes persuasion from coercion, deception, and incentivizing. Furthermore, the voluntariness condition follows by implication from this principle.

In section 2.4, I present the definition of PT proposed in this chapter, which has as its core this possibility of central route change, and test it by discussing a number of (purported) instances of PT.

Finally, in section 2.5, I first discuss which mental states are potential targets for change by persuasion (2.5.1), and I present three useful distinctions with regard to PT that enable the development of typologies of PT (2.5.2). Subsequently, I make some suggestions about the communicative nature of technological persuasion and the role designer intentions play; this because my definition states that technological persuasion is always an act of communication (2.5.3)

I conclude chapter two by noting that the proposed definition of PT is successful in the following respects: i) it specifies a principle or mechanism of persuasion, ii) it thereby enables a distinction between persuasion and other mechanisms of change, iii) it provides a conceptual unification of examples of PT, and iv) it has the desired narrower scope than the standard definition by viewing PT as essentially communicating technology.

2.2 The psychological processes underlying persuasion

In order to obtain knowledge of the psychological processes that underlie persuasion that will enable the formulation of an underlying principle or mechanism of persuasion, I will examine one prominent model of these processes, the elaboration likelihood model (ELM) of Petty and Cacioppo.[9] The ELM is a member of the big and growing family of so-called dual-

[9] In future research, I may also consider other influential models, e.g. the systematic heuristic model (SHM) of Chaiken and Eagly which is also a dual-process or dual-system model and which shares important characteristics with the ELM [Kruglanski & Thompson, 1999, pp. 87-88]. Of course, the ELM is only a model and one could argue that it is risky to tie the definition of persuasion to one specific model. However, even if the ELM model may be replaced by a better model, there is the growing body of empirical knowledge on persuasion that supports the ELM and considered the growing consensus in social psychology on the merits of the family of dual-process / dual-system, the model replacing the

process or dual-system models, which is specified to explaining persuasion. All dual-system models propose two modes of information processing [Strack & Deutsch, p221]: a slow, effortful, flexible and often conscious way of processing versus a fast, effortless, inflexible and often unconscious and automatic way of processing. Strack and Deutsch call these two modes the 'reflective' and 'impulsive' system. These two modes can be associated to respectively the central and the peripheral route of the ELM. As already noted, the ELM assumes as evident that persuasion essentially is communication, and I will follow the model in this respect[10]. By the term 'elaboration', the model means thinking about issue-relevant information. "Depending on the degree of elaboration, two types of persuasion processes can be engaged (one involving systematic thinking and the other involving cognitive shortcuts) – with different factors influencing persuasive outcomes depending on which process is activated" [O'Keefe, 2002, p 137].

This "dual-route but multi process theory" [Petty et al 1999, p157] distinguishes a *central*, thoughtful route to persuasion from a *peripheral*, low-thought route:

"Central route attitude changes refer to those that occur when people are both *motivated* and *able* to engage in relatively effortful information processing activity aimed at scrutinizing and uncovering the *central* merits of the issue or advocacy. Peripheral route attitude changes are characterized by low degrees of issue-relevant elaboration. Some peripheral route attitude changes are based on processes that differ primarily in *quantitative* ways from central route processes (e.g. elaborating few rather than many bits of issue-relevant information), but other peripheral route changes result from processes that are both less effortful and *qualitatively* different (i.e. doing something else than elaborating issue-relevant information [such as heuristic processing of simple cues, see Petty et. al., 2005, p86]) [Petty et . al. 1999, p157, first three emphases JS]".

As an example, consider and advertisement for an I-pod, which contains a photoimage of the ipod, and a person who is giving praise to the I-pod in the form of a text-balloon that contains five statements about the I-pod. According to the ELM, a person who is motivated to scrutinize this ad (e.g. because she wants to buy an I-pod) and who is able to do so, because of general capacities and specific knowlegde about what technical specifications one can expect for what price, will engage in elaboration. For this person, the ELM predict a central route change of attitudes. The perceived source expertise and the number of arguments (five) will not result in making her attitudes to the I-pod more positive, but

---

ELM model will be a member of the family as well, and I expect only minor consequences for my project of defining PT. It is even perfectly conceivable that in the end it becomes clear that there is only one system, which still facilitates two routes to persuasion. A last remark about this family of models: the term 'dual-system' seems more appropriate than 'dual-proces', because each system can harbor more than one psychological process.

[10] See section 2.5 below for a discussion of the relation between persuasion and communication.

argument strength on the contrary will do so. Another person, who is not motivated and able, will process the ad via the peripheral route and the ELM literature predicts that perceived source credibility and the relatively high number of five arguments will make his attitudes toward the I-pod more positive. He engages in using simple heuristics, which is chararacteristic of the peripheral route.

The ELM has a number of important characteristics, sometimes in the form of assumptions that are more or less confirmed by the empirical evidence. First, as becomes clear from the above citation and example, both motivation and ability influence the amount of thinking. Second, this amount of thinking is always located somewhere on the elaboration *continuum*, ranging from high (central route) to low (peripheral route); persuasion therefore involves not necessarily either high or low cognitive processing. Thirdly, the ELM assumes that people "want to have correct attitudes and beliefs, because these will normally prove to be most helpful in getting through life" [Petty et. al. 2005, p 83] Fourthly, and of importance to the study of persuasive technology, the ELM assumes that "…attitudes changed by high amounts of thinking are stronger than attitudes based on little thought. That is, such attitudes are more accessible, stable, resistant to counter messages and predictive of behavior", an assumption which seems to enjoy some empirical support [Petty & Brinõl, 2008, p 54]

In line with preceding persuasion research, the ELM distinguishes the following so-called persuasion variables: source, message, channel, recipient, and context [Petty et. al. 2005, p83]. Communication variables can be defined as "any aspect of the source (e.g. credibility, expertise, attractiveness), message (e.g. number of the arguments, strength of arguments), recipient (e.g. mood, ability, personality), or context[11](e.g. presence of distraction) that can vary in a given persuasion situation" and that can influence persuasion [Petty & Brinõl, 2008; subsequent discussion based on Petty & Brinõl, 2008]. These communication variables can influence the persuasion by affecting the following set of processes: the amount of thinking, the direction of thinking, structural features of thought, or whether information serves as a substantive argument or as a simple cue. Importantly, under different circumstances, the same communication variable can have different persuasive consequences, by affecting these processes differently.

As an illustration of this rather abstract statement of the ELM, I will briefly give some examples for each member of the above mentioned set of processes, from the research that Petty & Brinõl discuss. Source factors such as expertise, and attractiveness *serve as a peripheral cue* under conditions of low elaboration. "For example, when the personal relevance of a message was low, highly expert sources produced more persuasion than sources of low expertise regardless of the quality of the arguments they presented" [57].

[11] In the 2008 article, the persuasion variable 'channel' is absent and perhaps incorporated in 'context'.

Thus, the source is perceived as an expert and this serves to the recipient, without elaborate cognitive processing, as a simple heuristic cue. Under conditions that produce a high amount of thinking, source expertise *serves as argument* in favor or disfavor of the message.

Under conditions of moderate ability and motivation, the *amount of thinking* can be affected by, amongst other factors, the mood or the emotional state of the recipient. Typically, happiness leads to a decrease in the extent of message processing compared to sadness, which on the contrary leads to an increase of thinking. Under low or high elaboration likelihood conditions, the mood can have still other ways of influence.

An important *structural feature of thought* is the amount of confidence that people have in their thoughts that are generated as a result from a persuasive message. On the ELM, the favorability towards the issue of these thoughts is distinct to the confidence that recipients have in them. Under high elaboration conditions, source credibility increases this confidence [57].

Matching or tailoring the message to the recipient seems to *bias the direction of thinking* of that recipient under conditions of high elaboration, i.e. whether the recipient's thoughts are favorable or unfavorable to the persuasive message. In contrast, under conditions of low thinking, matching is more likely to *serve as a peripheral cue*. That is, if the message contains some cues that indicate matching with the recipient personal values, this can be enough to accept it.

It is important to note that according to the ELM, as can be inferred also from the examples, each communication variable can play a role both under conditions of low and of high elaboration likelihood. Thus, it is a misunderstanding to locate the good reasons (as relative to the recipient) contained in a persuasive attempt exclusively in the message. To illustrate, according to the ELM, the physical attractiveness of a person giving praise to a certain health product can under conditions of high amount of thinking potentially serve as an argument in favor of that product. Conversely, not all communication variables pertaining to the message are related to the goodness of its reasons; the number of arguments for instance, is a message factor, but has no bearing on the goodness of the reasons a message gives independent from the argument quality. In fact, if their quality is low, the empirical findings state a negative impact of a higher number of argument ('ten bad arguments is worse than five'). In the next section, I will investigate what help knowledge of these psychological processes that underlie persuasion can be to defining PT.[12]

[12] I want to put in one caveat. It seems that in contrast to other dual-mode theorists [Strack & Deutsch, 2004, p237-8], Petty et. al. eem not to incorporate the recent insight that attitudes can be implicit (related to the impulsive system) or explicit (more related to the reflective system). The attitudes formed by high amounts of thinking are most reasonably conceived of as explicit attitudes. However, according to recent research (Frank Verberne, personal communication), implicit attitudes are

2.3 Toward a positive characterization of 'persuasion'.

2.3.1 How the ELM illuminates the broadness of persuasion.

As was noted already, persuasion is a broad phenomenon and this broadness is also reflected in different dictionary definitions. One of them defines "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" [Merriam Webster Online Dictionary, 2011, 05-07-2011]. Another dictionary defines "to persuade" much broader as "to cause something, esp. by reasoning, urging or inducement; to induce to believe something" [Websters New World Dictionary, cited in Nickel, work in progress].

The ELM as briefly exposed above can account for the broad range of meanings denoted by these dictionary definitions. The first definition has a focus on mechanisms of persuasion in each of which a certain message plays a central role: in argument, entreaty, and expostulation a message is communicated. Nonetheless, they differ considerably in character, which can be explained by a different impact of the different persuasion variables (source, message, channel, recipient, and context), and a difference in the amount of cognitive elaboration that they receive. For example, it seems likely that in persuasion by argument, the amount of thinking is high, and the message, *serving as argument*, receives much attention and to the extent that the other persuasion variables are processed, they also (mainly) serve as argument. In cases of expostulation, it is conceivable that for example the context is such that the conditions (ability and motivation) are unfavorable to high amounts of thinking, leading to a major role of the mood of the recipient, processed as a heuristic cue. Importantly however, also in cases of expostulation, a message is communicated and if the recipient would have been able to engage in elaborate cognitive processing, it would have been open for her to assess the expostulation on its merits.

This kind of reasoning also applies to the first two mechanisms specified by the second dictionary definition: reasoning and urging. With regard to the third, 'inducing': this is such a broad category (including even coercion [Nickel, work in progress], in cases in which the inducement is too strong to resist) that it is difficult to apply the ELM model to the category as a whole. More generally, the ELM can account for the broadness of the phenomenon of persuasion by distinguishing countless aspects of source, message, channel, recipient, and context, all of which can be processed with a cognitive measure of elaboration ranging from high to low. In this way the model can illuminate appeal to

---

generally most predictive of behavior. The ELM however links these explicit attitudes strongly to behavior.

authorities, emotional exhortations, and some subtle forms of social influence and show how these can be elements of persuasive attempts.

2.3.2 The principle of persuasion and its difference with other mechanisms of change.

I would now like to show how the insights provided by the ELM contribute to a better conceptual understanding of persuasion and PT. My proposal is that every persuasive attempt is by definition an act of communication that in principle *can be* processed with high cognitive elaboration. That is, I define the core or the principle of persuasion therefore as follows:

Any persuasive attempt *is an act of communication* that enables *recipients, or* make it possible *for them, to assess their reasons for compliance with the target change by engaging in high amounts of issue-relevant thinking*.

In other words, the central route to persuasion must be open. For this route to be open, it is not enough that high amounts of thinking are possible. As I interpret the ELM, the central route consists of two elements: i) high elaboration ii) in order to assess the central merits of the issue.[13] Thus, the central route to persuasion is not only characterized by the quantity of conscious and reflective cognitive processing, but also by the aim of this elaboration: to arrive at a well-supported judgment of the persuasive message.

More needs to be said about the 'possibility' or 'enabling' clause of the definition, for it must be a real-world possibility, for real-world, fallible, and imperfectly rational human beings. Here again, the ELM provides ideas for specifying this condition. If the central route is open, users who are both motivated and able, *will actually* engage in high amounts of issue-relevant thinking; if they are both motivated and able, the only condition for actually taking the central route, is that route being open. So, if they *do* take the central route, this implies that it was open.

In order to understand persuasion, it is crucial to see that this definition of the principle of persuasion specifies a *condition* on persuasion, but not its *actual mechanisms that do the work of change*. The principle only states that the central route must be possible, but not that it must be the actual route to persuasion. The actual route depends on the

[13] "Central route attitude changes refer to those that occur when people are both motivated and able to engage in relatively extensive and effortful information processing activity aimed at scrutinizing and uncovering the central merits of the issue or advocacy" [Petty et. al. 1999, p 157].

persuasion variables in ways explained above; the same persuasive attempt may well be processed completely differently by two recipients.

This way of defining the underlying principle of persuasion has some virtues. It leaves intact the broadness of the phenomenon of persuasion by not specifying one definite actual mechanism of change; given the multiple psychological processes that potentially play a role in persuasion, this would be impossible. Nevertheless, it does provide a unification of this broad phenomenon by giving a necessary characteristic of each persuasive attempt, in terms of a route of change that must be actually possible. In this way, the principle can be made operative: it positively specifies, in terms of the ELM, what must be the case, in order for a change to count as persuasion. Fogg's voluntariness condition, on the contrary, does not by itself specify anything positively about the process of persuasion.

How does this principle of persuasion, in essence the possibility of central route change, enables the distinction between persuading, convincing, manipulating, and coercing? To start with convincing, it is natural to distinguish convincing and persuading by reference to the intent of the source. If someone wants to convince a person, she wants that person to change his beliefs on the basis of insight. From the perspective of the ELM, the source intention is to bring about central route change, and if someone is actually convinced, then this intention is realized. We could say that if an attempt to convince involves a clearly stated message, containing sufficient information and arguments relative to the issue at stake, which message is processed by an able and motivated recipient thus engaging in elaborate cognitive processing, in a context that also otherwise increases the elaboration likelihood, we have specified the paradigm case of convincing in terms of the ELM.

A neutral and technical meaning of the word manipulation is 'handling', 'manoeuvring' [Brenkert 2008, p155-6]. In the case of objects, this can be achieved by physical means, but in the case of humans this would amount to coercion. In order to bypass human resistance in handling them (thus treating the person as an object), the manipulator has to hide the fact that he is manipulating the person in some way. Therefore, the principle underlying manipulating is controlling persons in such way that their awareness of the very fact that they are manipulated is prevented. When this is done by giving people false or incomplete information, we speak of deception.

Importantly, manipulation is ruled out by the principle of persuasion. Manipulation, consisting in giving false or incomplete (in a specific sense) information and makes it impossible for recipients to assess their reasons for compliance. In case of giving false information this is evident, but in case of incomplete information, the issue is more subtle. The following distinction, that has to do with framing, seems crucial to avoid an interpretation of the principle of persuasion that implies a duty to give always complete and comprehensive information in persuasive attempts; for this clearly is not and need not always to be the case.

Consider two messages that contain exactly the same amount of information. The first (implicitly) presents this information as part of the complete set of information relevant to the issue, but the second suggests that the information is complete. In the first case, the recipient will probably avoid basing his assessment of the issue completely on the message, or at least he has the possibility to do so. In the second case, the recipient might base his assessment solely on the incomplete information of the message, in which case he is deceived and manipulated.

Stated very generally, manipulation can proceed in three distinct ways, all of which violate the principle of persuasion:

i) providing false or too incomplete information (deception; discussed in the previous section)

ii) providing no comprehensible information at all, but only some ambiguous cues that likely lead to the target change of the manipulator (subliminal advertizing?)

iii) providing information, but manipulating the context in such way that the central route is blocked and the peripheral route will, via some cues, likely lead to the target change of the manipulator.[14]

The underlying principle of coercion is that it involves physical necessity or such a strong incentive that this is psychologically impossible to resist acting upon this incentive. For example, you can either force someone in the desired direction by physically guiding him or by threatening him to kill his son. Coercing is ruled out by the principle of persuasion, because it implies that the central route to persuasion is blocked. In case of physical coercion this is evident and in case of psychological coercion, the recipient doesn't change his behavior because he comes to believe, relative from the perspective of her existing structure of beliefs, attitudes, and values that the target behavior is, in itself, the good thing to do, but because of a severe threat to something or someone he values.[15] The act of

[14] The following is an example of what I have in mind here. The issue of the extent to which people are aware and can be aware of attempts to manipulation is complex, which can be illustrated from the context of advertising. Both Brenkert 2008] and Wilson [2002] judge the frequent smoking in movies as manipulative. Of course, for example, an adolescent is aware of the fact that he sees someone smoking, but he is largely unaware of the way in which the frequent portraying of smoking in certain contexts, and the displaying of certain attitudes, creates his association of smoking with independence and rebellion. However, it could be argued that one has the opportunity to engage in thinking about how to evaluate the smoking of actors. It could be that such conscious reflection prevents association of smoking with independence and rebellion. So, even if the adolescent is unaware, it is open for him to be aware, were he motivated to do so. Perhaps the crucial question then becomes to what extent it is reasonable to expect such reflection from an adolescent.

[15] Here I have to think further, because in a sense, given the threat, the recipient has good reasons to comply with the target change, because a disincentive is always, in a specific sense, a reason. And, the recipient can engage in effortful thinking about the threat itself. However, in case of coercion he does not decide to comply with the target change as a result from high amounts of thinking about the central merits internal to this the target change, but he complies as a result of the strong external disincentive.

communication involved in coercion is only the channel of what does the real 'work' of change: the strong disincentive formed by the coercive threat.

In this section I first showed how the ELM perspective on persuasion matches nicely with the broadness of this phenomenon. Subsequently, I proposed to define a principle of persuasion as an act of communication that provides the *possibility* of central route change. This definition of the underlying principle of persuasion both matches with actual usage of the terms persuading and persuasion and also enables us to mark the distinctions between persuading, convincing, manipulating, and coercing. Therefore, in the next section I extend this way of specifying the core or the principle of persuasion to a conceptualization and definition of persuasive technology.

2.4 A definition of persuasive technology and its testing.

Based on a combination of all the, more or less explicit, insights gained so far I propose the following definition of persuasive technology. Persuasive technologies:

i) are intentionally designed to change (i.e. form, alter, or reinforce),
ii) through an act of communication,
iii) the behavior or mental state(s) leading to behavior of its users,
iv) while *enabling* users, or make it *possible* for them, to assess their reasons for compliance with the target change by engaging in high amounts of issue-relevant thinking (i.e., central route change is possible)

Compared to the standard definition by Fogg, this definition differs in the following respects. Persuasion is more narrowly defined as an act of communication[16]. Within the general category of 'change', specific types of change are distinguished, and 'attitudes' is replaced by the broader category of 'mental states that determine behavior'. All these differences will receive some attention in section 2.5. Most importantly, a principle or mechanism of persuasive change is specified: the possibility, provided through an act of communication, of central route change. In this positive characterization of persuasion, the voluntariness condition is implied and needs no explicit mention any more; the same is true for the fact that PT is neither coercive nor deceptive (manipulative).

[16] Although Fogg does not explicate his view on the relation between persuasion and communication, the context in which he presents his definition is about 'human-computer interaction'. In this interaction, messages and information is exchanges, so we can speak of at least a basic kind of communication. I take it that the far majority, if not all, of the examples that Fogg gives involve communication as the vehicle of persuasion.

In what follows, I will test this definition by discussing a number of concrete cases of potential PTs and in the course of this, the specific nature of *technological* persuasion will be illustrated.

- *Door handle*

The door handle does change our behavior and involves (until we use them in a habitual, automatic fashion) cognitive elaboration, but the reasons are inferred by us from the physical make up of the handle and the door. The handle doesn't communicate with us and therefore it is no PT.

- *Belief-or- behavioral-disposition- inducing-pill*

Although this pill is offered and taken in a persuasive context, the pill itself doesn't communicate and doesn't provide reasons for change by itself, and therefore, in line with our intuition, falls outside the scope of the definition. Its mechanism of change is of a biochemical nature.

- *Speedbump*

A speedbump is in the first place a coercive technology, because the mechanism by which a driver decides to reduce speed is the fear of damaging his car. This amounts to a coercive threat and thus to psychological coercion; the central route is blocked. The second reason why the speed bump is no PT is the fact that it doesn't communicate with the driver: it sends a message only in a metaphorical sense. Actually, it is us who infer the coercive threat from the physical make up of the speed bump[17]. If the speed bump would be replace by an electronic display that reads: "reduce your speed to 30 km/h, otherwise we will bring damage to your car!", it would be communicating technology, though still coercive and not persuasive.

- *Fasten your seat-belt blinking light and noise*

The blinking light and accompanying noise in your car that shines and sounds urges you to fasten your seat-belt. Because the lights have the shape of a person with seat-belts, this technology is literally communicative: it transfers the message that your seat-belts are not fastened. Now, in an average Western country, a driver could have (some or all of) four different reasons to fasten his seat-belts. First, he is aware of the safety aspects and these are reason enough for him to put on his seat-belts. Second, he is aware of the legal duty to

[17] In that sense, I find the 'script' metaphor of Martha Akrich (and Latour [1992] following Akrich) too suggestive. A script for a film or theater play is a literal, written communication. But speed bumps do not communicate in such literal sense.

fasten his seat-belt and this is reason enough for him to comply. Third, he is aware of the legal duty to fasten his seat-belt and also of the penalties for non-compliance and his fear of penalties is sufficient reason to comply. Fourth, the prospective to have to put up with the annoying sound all the way is reason enough to fasten the seat-belt.

In relation to the first three reasons, a light without sound, blinking only for five seconds would be sufficient for compliance. In the first two cases, this would lead to instances of technological persuasion, because central route change is possible. In the third case, depending on how much the driver fears penalties relative to the annoyance of driving with fastened seat-belts, we can speak of psychological coercion. However, the technology only reminds the user of the existing, external threat of legal penalties. It does not create those threats and therefore it is probably still a PT. The fourth reason to comply, on the contrary, *is* created by the PT. This case is complicated, because it is possible that reasons one and two lead the user to fasten his seat-belt, as soon as the PT 'reminds' him. Thus, it is possible that the central route to persuasion is taken. For users for which reasons 1-3 don't suffice, let's say after considerable issue-relevant thinking, compliance is secured by a coercive psychological threat, in this case provided by the technology itself. In other words, in this case taking the central route can lead to compliance, but not to refusal to comply, because then the technology will coerce. It is the possibility to act either way upon thinking about the central merits of the issue that defines the principle of persuasion.

The interesting feature of this case is that we have a technology that changes the behavior of some users already by persuasion but goes to greater lengths with other users (for whom persuasion is not enough for compliance) and in fact coerces them. The obvious drawback is the instability of the change in user behavior: as soon as he finds a way to silence the technology (e.g. by fixing the seat-belt under the seat)[18], he will be 'enjoying the real freedom of car-driving again'.

- *Ambient persuasion through light*

It has been shown that the energy consumption of people using a computer, who are therefore under a "cognitive load", i.e. they use their cognitive capacities for using the computer, can be influenced by using green or red background colors [Ham & Midden, 2010]. I interpret this as act of communication, though very implicit, through the symbolic meaning of the colors green and red, symbolizing approval and disapproval. According to the ELM, this background color will probably be processed via the peripheral route; in fact, this assumption is the guiding hypothesis in the development of this ambient PT. But

[18] The nice feature of making this PT really persuasive, e.g. by letting it stop after 30 seconds, is that the often occurring unintended ways people try to counteract such coercive means do not happen. See [Brey 2006] for an example of such side-effects of coercive technology.

importantly, because the light is visible and users know the symbolic meaning of the colors, the central route to persuasion is open as well and I therefore classify it as PT.

- *Subliminal "persuasion"*

Recent research [Ruijten et.al., manuscript] showed that by subliminal priming, it was possible to activate goals of people bypassing their conscious awareness. Importantly, this priming had a significant effect on subsequent behavior. According to the definition, this is no PT, because, although through priming short messages can be communicated, the central route to processing these messages is not open, because they are not available in the conscious awareness.

- *Influencing patient recovery by (day)light*

I can be brief on this example, because normal day light has no cognitive meaning in the way the colors red and green do, and thus, a technology stimulating patient recovery by employing day light, doesn't communicate and thus is no PT. It is also clear that it does not make sense to say that people engage in thinking about the light and as a consequence change behavior in healthy ways.

- *Glancing surface technology*

Imagine a technology that can make a surface glancing. Research has shown that in clean environments, the attitudes that are relevant to preventing people from littering become more activated.[19] Evidently, such technology would fail to be PT, because it doesn't communicate.

- *Weight loss website*

A weight loss website clearly communicates messages to users and also gives feedback to information given by users. This information exchange is explicit and allows, if reliable and complete enough, for central route change, and thus it is a PT. As is worth noting, these kinds of websites often in addition provide all kind of social support, which could be interpreted as providing incentives, and encouraging perceived behavioral control.

I think it is safe to conclude that, by means of providing a principle or mechanism of persuasive change, viz. the possibility of central route change through an act of communication, the proposed definition of PT is able to rule out cases which clearly are not instance of PT, to include many paradigm cases of PT and to shed at least some light on borderline cases. In short, the definition provides conceptual unification of cases that actually

[19] Cees Midden, personal communication, april 2011

are PTs and rules out the other cases in the way that should be expected from a proper definition.

2.5 Final issues

In this section, I will take up some definitional issues that were mentioned while introducing the defintion of PT in the previous section. First, I will pay attention to the question what specific mental states are target of (technological) persuasion (2.5.1). Second, I will provide three typologies of PTs (2.5.2). Finally, I will discuss the relation between persuasion, communication and designer intentions as relevant to the definition of PT (2.5.3)

2.5.1 What mental states can be changed by persuasive technology?

The standard definition identifies two possible objects of change: attitudes and behavior. However, by definition, persuasion never directly brings about a change in behavior, but always indirectly via changing one of the mental states of the human mind that are determinants of behavior. Attitudes form only one class of these determinants and furthermore, the attitude-behavior relation can be complex.

Attitudes can be defined as a person's general evaluation of an object (persons, institutions, policies, etc.) [O'Keefe 2002, p6; Ajzen and Fishbein 2000, p3]. Apart from changing valence and extremity of attitudes (its content so to say), persuasion can also be targeted to the following changes with regard to attitudes [O'Keefe 2002, pp. 20-25]. Persuasive efforts can be directed at enhancing the perceived relevance of an attitude in a certain situation, by pointing out the relation of the attitude to a certain behavior (attitude activation[20]). This attitude activation can be strengthened by encouraging the recipient to anticipate how she will feel upon performing the behavior. Accessibility of attitudes, confidence with which attitudes are held, strength of attitudes, and beliefs about the object of the attitude are further targets of persuasion.

Although in large strands of persuasion research persuasion seems to be identified with attitude change, typically no reason is given for this identification. And, it is difficult to see which reasons could be given, for clearly more mental states exists that determine behavior and that are for that reason potential targets of persuasion. Behavior cannot be directly changed by means of persuasion and if the concept of persuasion includes behavior change, then the burden of proof lies with the one who proposes to include attitudes into the

[20] As evidenced by numerous experimental studies in the last two decades, attitude activation can occur automatically and unconsciously [see also Ajzen and Fishbein 2000]. Following from the definition proposed in this paper, the important question is whether recipients of PT are potentially in control of this automatic and unconscious attitude activation.

concept but not the other mental determinants of behavior[21]. According to the 'Theory of Planned Behavior' [Ajzen and Fishbein 2000, p17], together with the attitude toward the behavior in question, also 'perceived behavioral control' and 'normative beliefs' (beliefs about the normative expectations of significant others[22]) determine behavior. A moment of reflection will learn that our everyday communicative and persuasive efforts regularly target exactly these two additional factors.

2.5.2 Three typologies of persuasive technologies.

If the concept of persuasion is defined so broad as in the definition of PT and as in the previous section, it is helpful to make some distinctions within the concept in order to facilitate the study of and ethical reflection on PTs. The following three typologies will prove helpful in general, as will be visible from the use of them in the next sections.

i) Macrosuasion versus microsuasion.

Fogg [2003, pp. 17-20, 247-8] uses the term 'macrosuasion' for persuasion by technological artifacts for which the designers intention to persuade and motivate users forms the 'sole reason' for their existence. Examples are the weight-loss website and the 'fasten-your-seatbelt-blinking-lights' discusses above (section 4).

Microsuasion refers to artifacts which overall goal is not to persuade, but that include 'smaller persuasive elements'. Fogg gives the example of persuasion- and motivation strategies built video games, in order to support the achievement of the overall goal, according to Fogg in most cases entertainment. More generally, Fogg discerns a trend to include more microsuasion into products with the aim to improve its functionality.

This micro-macro distinction in persuasion does not coincide with a certain way of making a parts-whole distinction within artifacts. An eco-feedback dashboard is part of the artifact 'car', whose overall goal is something like 'providing a means of transport'. The persuasion built into the eco-feedback dashboard is not intended to support this overall goal of 'transport' as such, but to reach that goal in a more sustainable way. Persuasion is therefore the sole reason of existence of the eco-feedback dashboard, which we can see as an artifact in itself, that is part of the encompassing artifact 'car'.

ii) Kinds of "change".

[21] Zimbardo and Leippe [1991, cited in Fogg 2003, p20] include in the concept a person's "behaviors, feelings, or thoughts about an issue, object, or action". The category 'thoughts' is too unspecific to be useful, but I endorse the broadness of this definition.

[22] The persuasive effect of the I-cat, a robot used at the HTI department of the TU/e . [see for exampleVossen et. al. 2010] can be partly attributed to activating these normative beliefs

The standard definition speaks of “change of behavior, attitudes, or both”. ‘Change’ is a very broad concept however and it is conceivable that one who uses this concept has in mind a more specific type of change. I have the impression that some people who speak of ‘changing attitudes’ have in mind the replacement of existing attitudes with new ones[23]. In any case, the broad category of change includes more types than replacement. Some changes related to attitudes described in the previous section do not involve a replacement of the attitude for example changes in strength, valence, and extremity. Oinas-Kukkonen [2010, (see section 2.1)] provides a nice typology of changes: formation, alteration, and reinforcing. In order to prevent that every potential user of the definition proposed in this paper will think and argue with her own more specific type of change in mind, I will not use the term ‘change’ in the definition, but instead the typology of Oinas-Kukkonen.

iii) Attitude change versus promoting attitude-behavior consistency.

A useful distinction runs between PTs that are primarily designed to bring about attitude change (formation, alteration, or reinforcing) and PTs that have the attitude-behavior consistency as their primary persuasive intent. That is, these latter PTs aim to ensure that people act in ways that follow from their attitudes. To recall from section 2.5.1, attitude-behavior consistency can be promoted by enhancing the perceived relevance of the attitude (activation), and the perceived behavioral control, and by activating normative beliefs.

The rationale of this distinction between these two types of PTs might lie in the fact that attitude change, and especially the formation of temporarily stable new attitudes requires higher amounts of information, arguments, and elaboration of these. It is typically the kind of change in which philosophers have a special interest. As can be inferred from the instances of PT discussed in section 2.4, most PTs are of the second type. This can be explained by the fact that the use-context of many PTs do not allow for extensive elaboration of extensive persuasive messages. Promoting attitude-behavior consistency on the contrary often needs less arguments and less elaboration. In some cases, norm-activation is simply sufficient (e.g. the blinking lights that remind you to put on the seatbelts). In other cases, the perceived behavioral control is enhanced by simple feedback mechanisms.

Though useful, the distinction is not at all absolute. A given PT can be designed with the persuasive intent both to change attitudes and to promote attitude-behavior consistency. Certain weight-loss websites will both encourage healthy attitudes toward food and try to motivate users to act on those attitudes. PTs of the second type may still change attitudes. An eco-feedback dashboard in a car that gives feedback on fuel use may enable driver A to

[23] Although it is not evident, Fogg might conceive of attitude change in this way, for he writes “Other scholars expand persuasion beyond the idea of “changing”; persuasion includes shaping and reinforcing” [2003, p 20, footnote 1]. However, evidently, shaping and reinforcing are specific types of change.

act on his positive attitude toward eco-driving. Driver B who initially has no such positive attitude, may form one by monitoring the difference in fuel-use between his normal vigorous driving style and eco-driving. Furthermore, psychologists point out that the attitude-behavior relation is also causally bidirectional: our behavior can also have an effect on our attitudes and other mental states [Strack and Deutsch, 2004]. Our actions can shape our self-perception [Cialdini, 2007]. Finally, according to the ELM, the way in which attitudes are formed also bears on attitude-behavior consistency, where attitudes formed under conditions of high elaboration are more predictive of behavior [Petty& Brinõl, 2008, p54].

The distinction as presented here is descriptive, serving as an analytic tool. It is not meant to evaluate one type as morally better or ethically more acceptable as the other, because such an evaluation is not at all immediately evident. Though attitude change might be conceived of as more central to a person´s self, as long as persuasion *is in fact persuasion*, i.e., allows for change based on high amounts of thinking about the central merits of the issue. On the other hand, it is too simple to say that PT that helps people to act on attitudes they already have is for that reason ethically laudable. Consider only the existence of, for example, racist attitudes.[24]

2.5.3 Persuasion, communication, and designer intentions.

In the definition I propose, persuasion is seen as a subclass of communication: not every communication is persuasive, but every persuasion is a form of communication (although of a very elementary form in many instances of PT). Furthermore, PT is defined by reference to designer intentions. In this paragraph, I want to make some comments on the relation between these two parts of the definition, though issues are complicated here and I do only make some tentative suggestions.

Intentionality plays a role in both persuasion and communication, but a different one. Persuasion involves the intention on the part of the persuader to change mental states that determine behavior, often with an underlying aim to change behavior in a specific way. Communication is, like persuasion, notoriously difficult to define [Dance, 1970]. According to philosophy of language approaches, the intention behind communication is to present information. Some communication theorists will insist on including the uptake of the information in the intention of the communicator. In this way, the intention involved in

[24] Racist attitudes are the example by means of which Wilson [2002] explains the difference between implicit and explicit attitudes. People may if asked sincerely report attitudes that are friendly to afro-americans, while clever experiments reveal their prejudicial implicit unconscious attitudes. It seems to me that the issue of attitude activation by PTs is a major ethical issue, precisely because of this complicated nature of attitudes. PT may affect both types of attitudes and the ethical evaluation might well differ, for it is the question whether activation of implicit attitudes by PT can be under control of the recipient.

communication comes closer to the intention involved in persuasion. Even so, persuasive intentions are more clearly directed to changing behavior or the mental states that determine behavior.[25]

Such reference to human (designer) intentions is more complicated in making the distinction between technology that merely communicates with its users and technology that persuades its users. I will assume that technology can communicate in the first place[26]. But even on this assumption, different intentions behind communicative and persuasive technology might make no difference on what actually happens when users use these technological artifacts. Consider the following examples. At first sight, a simple speedometer might not look like a PT, whereas the eco-feedback display can be called a paradigm example of a PT, at least it is often alluded to this way [see for instance Meschtscherjakov 2009, Spahn 2011]. However, on the proposed definition they are not at all so different: both involve (very elementary) communication, and allow for central route change. Why then might the eco-feedback display appear so persuasive compared to the speedometer? It seems to me that the explanation is the appearance of the presence versus lack of a *clearly visible persuasive intent*.

The origin of this appearance, I hypothesize, lies in the fact that the eco-feedback display is relatively new for us. The previous situation, which especially older drivers will be familiar with is the absence of eco-feedback. In the former standard situation, sustainability of energy was no issue and driver behavior was not influenced by it. Then, we became aware that our behavior should change, relative to this standard situation: we should be *persuaded* to change our behavior. In general, persuasive technology seems to involve a certain standard situation, which has some undesirable characteristic that have to do with the way (some) humans behave. This motivates designers to develop a technology with the explicit intention to change the behavior of these humans relative to the standard situation.

So, it seems to me that the prominent persuasive intent of the eco-feedback display makes it such a clear example of a PT. It is however an easy exercises to invent an analogous history, if it is not the true history, of the development of the speedometer. And if we look at speedometers with small red zones indicating the relevant maximum speeds (norm activation), these speedometers appear already somewhat more persuasive, because these red zones implicate an evaluative ought: 'you ought not drive faster than this speed'..[27]

[25] Anthonie Meijers suggested making the distinction between communication and persuasion from the perspective of speech-act theory. Communication and persuasion differ in their perlocutions.
[26] See Spahn [2011] for a discussion and defense of this assumption.
[27] It seems to me that there is some intuition behind that persuasion really involves a big and major change in existing attitude system because of new reasons and information from an external source. But, sometimes persuasion consist in just pointing out to someone the consequences of what he already believes [Nickel: Socratic influence] and sometimes, a simple factual belief is enough to result in major behavior changes.

By way of thought experiment, it seems perfectly conceivable to imagine two technically identical artifacts (e.g. a version of the speedometer), while the first is intentionally designed with an explicit persuasive intent, but the second is not. Is then the first a PT and the second not? The affirmative answer to this question has the unwelcome effect of making what a technology is too much dependent on the mind and intentions of designers and employers and too little on what happens in the mind of users when they are 'alone' with the artifact. A technological artifact that is designed to be persuasive can fail to be so and a technological artifact that is not explicitly designed to persuade can be change behavior or its mental determinants in unforeseen ways. Nickel [work in progress] speaks of "persuasive technology" versus "technology that persuades". If we on the contrary don't let designer intentions make the distinction between communicating technology and PT, then I see no way to distinguish between these two classes of technologies, while they certainly appear to be distinct, though the border between them will probably be diffuse. I will leave these questions to future research[28].

[28] Incorporation of the philosophy of technological artifact function might be helpful here.

3 Toward a framework for an ethics of PT

In the previous chapter, PT was defined by essentially two key-elements: it is communicative technology and it allows for change central route change (i.e. *enables* users, or makes it *possible* for users, to assess their reasons for compliance with the target change by engaging in high amounts of issue-relevant thinking). From this conceptualization of PT, it followed that PT is neither coercive nor manipulative, but allows for voluntary change. It could be wondered why such type of technology aimed at voluntary change asks for ethical reflection at all. However, even if PTs rely by definition on methods that respect the freedom of its users, for any given instance of a purported PT it has to be assessed what its *actual* voluntariness for change is and thus whether it is indeed a genuinely persuasive technology and not a (slightly) coercive or manipulative technology. This perhaps suggests that 'persuasion' and 'persuasive technology' are thick concepts: they have descriptive content, but at the same time, using them implies a moral value judgment.

Apart from methods of technological persuasion, also its target change, the final aim served by that target change and the *actual* outcomes deserve ethical reflection. For, the target change might be unethical in itself. Or, it may imply a distributive injustice to choose a certain audience as recipient of the technological persuasion. And even if they are free to remain unpersuaded, it might be conceivable that in some cases, even the attempt of technological persuasion may be a matter of not respecting their autonomy. For the final aim, the question can be asked whether it is most appropriately served by persuasion and PT or instead by education[29], convincing, or by coercion. Finally, because the outcomes depend on the success of persuasion, designers should consider the possibility that the target change is not brought about and the final aim is not reached or served. In addition to not contributing to the final end, other unintended outcomes may obtain. One possibility: in cases of unsuccessful persuasion, the PT might become unsafe if functionality depends on successful persuasion.

Most of the ethical questions just raised will receive a more detailed treatment in this chapter. In 3.1 a critical discussion will be given of four existing (attempts of) ethical frameworks. In section 3.2, using insights from all these existing ethical reflections on PT, I will give a start to developing a framework based on the distinction between methods, target change, and final end, outcomes (e.g. other than target change, serving other final ends or having other than intended outcomes (including safety and responsibility, e.g. when

[29] From the perspective of the previous chapter, education aims at central route changes, with the long term goal of raising autonomous citizen. Persuasion, although it must enable central route change, it is typically aimed at change of behavior, regardless of which route to persuasion is taken. Of course, the educational process contains a good deal of persuasion as a means.

functionality depends on successful persuasion). Because the ELM perspective of the previous chapter applies mainly to methods of persuasion, the focus of this section will correspondingly lie on ethical aspects of technological methods of persuasion. In section 3.3, I focus on an important aspect of the moral evaluation of PTs: the question as to what extent technological persuasion forms a threat to personal autonomy.

3.1 Discussion of existing frameworks

3.1.1 Berdichevsky & Neuenschwander: the golden rule of persuasion

In their pioneering article "*Toward an Ethics of Persuasive Technology*" [1999], Daniel Berdichevksy and Erik Neuenschwander (hereafter B&N) make the conceptually and ethically important distinction between designer motivations, methods and outcomes of technological persuasion[30]. Based on this distinction, they propose the framework for an ethics of PT which is given in figure 1 below.

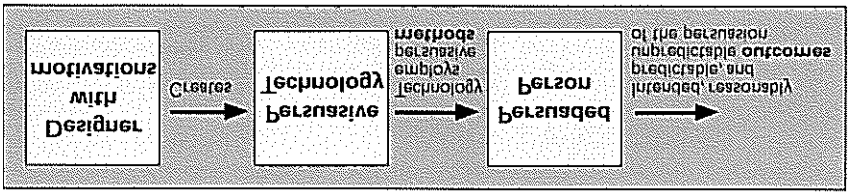

Figure 1. Berdichevsky & Neuenschwander's framework for the ethics of PTs.

I will explain this framework by using an example of the authors [55]. Suppose, three designers share the same persuasive intent to build a PT that persuades a stranger to eat more fruit and vegetables. Still, their motivations can differ, for example the first may be motivated to increase the health of people, the second to gain economical profit, and the third "by a secret hope the stranger will eat a bad fruit and become sick to the stomach". Clearly, the persuasive intent can be the same, but the motivations behind the intent differ, and do so in an ethically important way. With regard to the methods of technological persuasion, they note that these are the same methods as used by human persuaders, but embed these "…in a new and compelling context" [B&N, 1999, p 55]. The outcome of the technological persuasion is defined by the authors as "what the persuaded person is persuaded to do or think". An allergic reaction of the stranger, caused by eating a kumquat

[30] This distinction runs parallel to the distinction between intentions, methods, and outcomes of Fogg [2003, 220]. This can be slightly confusing, because Berdichevsky and Neuenschwander also use the terminology of 'persuasive intent', as distinct from motivations. (Berdichevsky and Neuenschwander worked in the lab of Fogg.)

would be an outcome unintended by the first designer of the example and intended by the third.

As is clear from this example, the concept of "outcome" is ambiguous between relating to the persuasive intent on the one hand and the realizations of the underlying (designer) motivations on the other hand. For, what the persuaded person is persuaded to do or think in this case is 'eating the kumquat', which means that the persuasive intent of all three designers is realized. However, the result of this realization of this persuasive intent is the allergic reaction, which means that only the third designer reaches the aim that motivated him. In addition, B&N use the term 'outcomes' also for unintended side-effects of persuasion. In the ethical framework developed in the next section (3.2) I will for this reason distinguish between 'target change' (in B&N's terminology: 'persuasive intent' and if realized, 'outcome') and final end, that is supposed to be served or fostered by the target change.

The final ethical evaluation of a PT and the ascription of moral responsibility depend on asking questions on three levels: whether the outcome was i) intended or not, ii) reasonably predictable or not, and iii) ethical or unethical. [see B&N's figure 5 on p 55] So, in the present example, the first designer clearly has an ethical goal and the outcome, the stranger suffering from an allergic reaction due to eating the kumquat, is unintended. His moral responsibility therefore depends on the question to what extent the allergic reaction was reasonably predictable. The third designer is morally responsible and at fault already by his evil intention alone.

In the B&N framework, the issue of the predictability of an outcome plays a role only in the case of *un*intended outcomes. However, this issue plays an equally important role in case of intended outcomes. The mere fact that an outcome is brought about by technological persuasion in the way that was intended, doesn't imply that it was reasonably predicted. Perhaps the chances of persuasive success were (or reasonably should have been) predicted as only 50 % and persuasive failure would have caused severe damage to the user or other negative consequences. In that case, the designer cannot be hold morally praiseworthy merely because it *happens to be* that the outcome is as was intended, and ethical; it must also be likely. Although not explicitly taken up in their framework, B&N will probably agree with this line of reasoning. For, from several of their examples it is clear that B&N endorse the view that designers have an important duty to "anticipate unexpected outcomes" and they are "…responsible for all reasonably predictable outcomes of their persuasive methods …[which]…requires significant user testing and holistic forward thinking on the part of designers" [B&N, 1999, p57]. It is reasonable to suppose that this duty automatically encompass predicting the intended outcome *on good grounds*.

B&N develop eight “ethical principles of persuasive design”[52], from which some are taken up in the present discussion. The eight’ principle is their final “golden rule” of persuasion:

“The creators of a PT should never seek to persuade anyone of something they themselves would not consent to be persuaded of” [58].

Background of this golden rule is the Rawlsian idea that you only can agree to be treated in ways that you would consent to from behind the “veil of ignorance [Rawls, 1999/1971, pp.102-160]]. This golden rules goes beyond identifying motivations, methods, and outcomes as aspects of ethical reflection and evaluation, because the golden rule gives material answers to the question of how an instance of PT should be ethically assessed, instead of specifying aspects of a procedural answer to this question. However, the golden rule will not always give the correct answers, because it may be either too strict or too permissive. B&N [57] themselves give the example of the smoking father that doesn’t want to give up his addiction, so, he would not consent to be persuaded to quit smoking. Still, the father might want to design a technology that persuades his daughter to quit smoking. Whereas in this example the golden rule turns out to be too restrictive, Spahn [2011] gives an example in which the rule is too permissive: a racist that builds a PT that tries to persuade people of racist idea’s which he holds himself and consents to be persuaded of.

In answer to the limitation of their golden rule, B&N point to the other principles they developed. However, the material guidance these principles supply is very limited. Principles 1-3 state with respect to outcomes and motivations that these should not be “deemed unethical” in a human-human persuasion context and furthermore with respect to outcomes that designers are morally responsible for all reasonably predictable outcomes. Clearly, this only emphasizes the importance of outcomes and motivations for ethical reflection, but what we still need to know is *why* a certain outcome is “deemed unethical”, or according to *whom*, or according to *which ethical theory* (in this latter case we also need to know why this ethical theory should apply to the ethical assessment of PTs).

This lack of material substance is very understandable given the current state of the field of applied ethics in which the facts of the fundamental moral pluralism in most Western societies propel ethicists into developing procedures for reaching actual consensus in technology assessment settings. From that perspective, B&N made a very important and helpful contribution to the ethical assessment of PTs.

3.1.2 Fogg: methods of technological persuasion and stakeholders analysis.

Fogg's approach to "the ethics of persuasive technology" [2003, pp. 211-239; see also pp. 7-11; page numbers in this section all refer to this work] is very similar to the approach of B&N, but in addition gives extensive attention to the ethical concerns which are specific to *technological* persuasion and discusses the method of stakeholder analysis for "systematically examining the ethics of any persuasive technology product" [233].

Human-human persuasion can involve very symmetrical relations but also considerably asymmetrical ones. For example, consumers sometimes feel bewildered after buying, from a meticulously skilled salesperson, a product that they don't need and that they don't want [Cialdini, 2007]. In technological persuasion, some of the sources of potential asymmetry between human persuadee and PT are similar (e.g. PTs that leverage social influence), but others are unique.

Fogg discusses several of those sources of asymmetry that are specific for PTs[31], which can be grouped together as follows:

i) PTs control the interaction. The way in which the PTs are programmed by their designers determines and *limits the interaction possibilities* that human users have. 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. The currently existing PTs on the contrary are very limited in their capacities for two-directional interaction. Related, PTs can also be *proactively persistent*, because unlike humans, they don't get tired, embarrassed, or start feeling guilty or uneasy. PTs can continue its persuasive attempt until the user gives in, in a moment of weakness or in a moment of unawareness. Another asymmetry between human user an PTs arises from the fact that PTs are (not yet) able to have emotional interaction with users, whereas it can show (programmed) emotions, which can be a powerful means of persuading humans.

ii) PTs (being or embedding computers) generally have great capacities. Computers can "*store, access, and manipulate huge volumes of data*"[8], which enables them to provide the right piece of information on the right place and time in order to persuade. Based on information on buying behavior of a certain customer and all their other customers, commercial websites like Amazon.com can make suggestions that are tailor

[31] Fogg speaks of computers, which form his focus. However, any PT is by the definition proposed in chapter two a communicating technology and therefore will at least contain a mini-computer. In this section I combine and merge most of the sources that Fogg mentions in his introduction and sources from his chapter on the ethics of PTs. Italics refer to Foggs original classification. Fogg gives his discussion not in terms of asymmetry but in terms of "advantages PTs have over human persuaders" an in terms of "unique ethical concerns related to PT".

made[32]. Beyond this, by gaining knowledge of which persuasive strategies are most effective for specific customers, and by even sharing this knowledge with similar commercial websites, PTs enable so-called persuasive profiling [Kaptein and Eckles, 2010]. Once successful persuasive strategies are developed, their application potential can be multiplied with the help of PTs: *computers can scale*. On the basis of their great processing capacities, computers can also *use many modalities*: audio, video, text, graphics, animations, and hyperlinks can be used and combined in order to enhance and tailor persuasion.

iii) PTs can have a wider range of access. Many PTs can go where humans cannot go and thus *can be ubiquitous*, e.g. in the bathroom or in the car, and this enables PTs to persuade at the right moment and place. Also, *PTs allow, in principle, for anonymity* and bypass the social barrier that exists for certain subjects in human-human persuasion.

Al the factors mentioned above can contribute to an asymmetrical PT-human relation. If this asymmetry is perceived by the user, they might experience the PT as a powerful persuader or even coercive technology, and consequently show reactance [Roubroeks et. al., 2010]. Think of registration programs that don't proceed any further unless to deliver certain information. These aspects of PTs thus point to the ethical need to assess whether an instance of PT indeed is persuasive or might better be called coercive. If the asymmetry reveals itself in a certain PT being extremely effective, without users feeling coerced, it might be the case that it is manipulative, which also calls for ethical reflection on methods[33].

Apart from methods, according to Fogg also intentions (functionally equivalent to B&N's 'designer motivations') and outcomes deserve ethical attention. This can be done by a stakeholder analysis, in which the following steps should be taken: i) list all stakeholders and ii) what each has to gain, and iii) has to lose. Subsequently iv) evaluate which stakeholder has the most to gain, and v) which has the most to lose, and, finally, vi) "determine ethics by examining gains and losses in terms of values" [234]. In doing so, vii) the person that carries out the stakeholder analyses should be sensitive to which values and assumptions serve as input.

Whereas B&N give some substantial guidance by their golden rule of persuasion, this method of stakeholder analysis provides only a valuable method or procedural approach to the ethics of PT. However, this method (and related methods as well) leaves a plethora of issues undetermined. To mention just a few: which parties should carry out the analysis and why? What are the criteria that determine who is a stakeholder and who is not? When are

[32] From the literature on persuasion, it is known that tailoring is a powerful persuasion strategy [Petty & Brinol, 2008, p58] This strategy can of course also be used by human persuaders, but it is effortful and that is exactly the reason why computers can utilize its persuasive potential so well.
[33] See Nickel & Spahn [submitted] for guidelines to 'PT design for symmetry'.

the lists of losses and gains complete? Step 6 might imply a broadly conseqentialist framework, but how can the relative weight of values be determined? [Van de Poel, 2009] And perhaps some values, for example autonomy, might have an overriding character in certain contexts. Although it is very much preferable that every designer of PTs carries out a stakeholder analysis to sharpen his awareness of ethical issues, this analysis does not in itself enable the designer to decide on the moral acceptability of a given PT. Nevertheless, Fogg offers some substantial insight by giving concise ethical evaluations of several instances of PTs.

3.1.3 Verbeek: Expanding the B&N framework with mediation analysis.

In his paper on the ethics of PT [Verbeek, unpublished], Verbeek develops an expansion of the framework of B&N (see section 3.1.1) with his theory of technological mediation. This theory emphasizes that artifacts are not merely functional instruments, but "….active mediators in the relations between humans and reality" [3]. In addition to serving as tools that humans use to reach their goals, firstly technologies "…help to shape how reality can be present for human beings, by mediating human *perception* and *interpretation*; second, technologies help to shape how humans are present in reality, by mediating human *action* and *practices*"[p3].

The next step Verbeek takes is to subsume persuasion under the umbrella of mediation; he views technological persuasion as a specific form of technological mediation[34]. Beyond this, the persuasive function "can also have a mediating effect *itself*". [p4]. As an illustration of these two forms of mediation Verbeek uses the nice example of the FoodPhone, a PT based on mobile telecommunication that is designed to persuade users to develop a healthy eating pattern, likely bringing about a loss in weight. The persuasive function can, according to Verbeek, be seen as a form of mediation, because it shapes it users' interpretation of what they are eating. However, apart from the desired outcome, the persuasive function can also lead to additional forms mediation that are less desirable, e.g. eating can become a stressful activity, or, the fact that people have to take pictures of their food can have a negative influence on the social atmosphere and change the practice of eating.

Whereas B&N pay systematic attention to both intended and unintended outcomes of technological persuasion, Verbeek adds to this the outcome of *all mediations* that arise in

[34] See below for a critical discussion of this view. See Waelbers [2007] for a critique of the way Verbeeks' concept of mediation serves as an umbrella term.

the use of technology and not only those linked to the persuasion. This leads him to the identification of "three points of application for moral reflection" [pp. 7-8]:[35]

i) Intended persuasions: can they be morally justified?

ii) Forms of mediation (including methods of persuasion): are methods and implicit forms of mediation morally acceptable?

iii) Outcomes of mediation: "can the consequences of the persuasive and otherwise mediating role of the technology be morally justified?"

Verbeek discusses how a stakeholder analysis could be done for each of these three points of application. Especially with regard to the third point, the outcomes of mediation, the tool of mediation analysis needs to be used: "with the help of moral imagination, an inventory has to be made of all possible mediating roles of the technology in both human experiences and human actions" [10].

On my interpretation, Verbeeks's two forms of mediation linked to technological persuasion (i.e. technological persuasion as a specific form of mediation, and the mediating effect of the persuasive function itself) run completely parallel to B&Ns' distinction between intended and unintended outcomes of persuasion. The valuable addition then, is to take into account also the non-persuasive forms of mediation that always arise when technologies are used. I see it as a regrettable shortcoming however, that Verbeek does not conceptually clarify in what way persuasion is a specific form of mediation, as he claims. He only distinguishes persuasion from "forcing" and "seducing" and gives an example of technological persuasion that employ a feedback mechanism, but nowhere gives a conceptual clarification or even definition of PT.

This lack of conceptual preciseness inhibits ethical reflection, as becomes clear from Verbeeks' discussion of the legitimacy of PTs as a class of technology. He lumps PT into the class of "behavior-influencing technology" and then proceed to discuss the questions of user freedom. However, the "material inhibition imposed by a speed bump" [13] limits user freedom in a completely different way as for example an EcoFeedback device does: the first is clearly coercive, the second persuasive (see section 2.4 above). Given Verbeeks emphasis on 'materializing morality' and on the fact that artifacts can "mediate action *as material thing"* [Verbeek 2006, p368, emphasis added], this lumping may come as no surprise.

However, whereas Verbeek tries to show that engineers are "doing ethics by other means" and are "materializing morality" [Verbeek 2006, p361], it could be argued that

[35] This is basically the same three-part distinction of B&N and Fogg between intentions, methods, and outcomes. As was noted above with respect to B&N, here as well the notion of "outcome" is ambiguous between relating to what I name "target change" and "final end".

designing persuasive technologies in fact amounts to *de*materializing morality. If PT is conceived as allowing, trough an act of communication, for central route change, then the mediating effect bound to the persuasive function is of a *communicative nature* instead of material. Because users are enabled to base their change in behavior and or attitude on their thinking about the issue, the persuasion-mediation is made explicit and dependent on user cooperation. So, the morality is by design located in the human subject and not in the material object.[36]

3.1.4 Spahn: discourse ethics applied to PT.

Spahn [2011, forthcoming[37]] develops ethical guidelines for use and design of PTs by applying two elements of discourse ethics to PT: the idea that communication is an inherently normative activity, and the distinction between 'communicative' and 'strategic rationality'. His focus is on the methods of persuasion and not so much on the issue of the moral acceptability of the target change and final aim. Because persuasion is an act of communication, every persuasive attempt can be seen as a speech act. Jürgen Habermas argued that in a speech act (Austin), the speaker makes four implicit validity claims, which are by Spahn applied to PT in the following way:

i) Comprehensibility: it must be possible to understand the utterance. For PT this means for example that if a PT employs evaluative feedback as an element of persuasion, the meaning of these often simple cues (e.g a red or green light) should be clear to the user.

ii) Truth: the information given by the PT should be true.

iii) Truthfulness: because a PT has no mental states like humans, it cannot be truthful in the strict sense. Therefore, in case of PTs (interpreted as the persuader), truthfulness can be equated with the reliability and accuracy of the mechanisms by which a PT exchange information with the users.

iv) Appropriateness: one interpretation is the question whether the employment of a PT is an appropriate means to a given end. To answer this question, additional ethical theory is needed.

[36] Note that I do not argue against the idea of materializing morality as such; I only claim that in some respects, PTs do not fit into this approach. A careful discussion of the "politics of PTs" should clarify what ends should be pursued by designing PTs.

[37] Similar conclusions (though not based on discourse ethics) are reached by [Davis, 2010], and [Baker & Martinson, 2001]

These four criteria can thus be applied in a meaningful and insight giving way to PT, but they "are still on a very abstract level (as they can be applied to all types of communication) and may thus not lead to more substantial rules for what is and what is not allowed in the specific contexts of *persuasion*". Therefore, Spahn turns to the Habermasian distinction between strategic rationality (finding the adequate means for an end) and communicative rationality (aimed at a common search for insight on the basis of good arguments). Now, 'persuasive rationality' "seems to fall exactly in the middle between" strategic and communicative rationality. Because from discourse ethics, communicative rationality is the ethical ideal, Spahn proposes three guidelines for the design of PTs that should bring persuasive rationality as close as possible to communicative rationality. First, "persuasion should be based on prior (real or counterfactual) consent".[38] Second, "ideally the aim of persuasion should be to end the persuasion". Third, "persuasion should grant as much autonomy as possible to the user", in order to approach the ideal of communicative rationality which is centered around autonomy.

It seems to me that Spahns' approach can be summarized by the following slogan: "design PTs as much as possible as *convincing* technologies". As said, this approach follows from the discourse ethics ideal of communicative rationality. The four normative criteria on every communication act help to shape PTs into convincing technologies. However, Spahn himself acknowledges that it depends on the situation whether convincing, manipulating, or coercion is the morally right method to change a persons' behavior. By implication, there are also situations in which persuasion is the proper method. Therefore, although Spahn leaves this question as to *when* to use PT for future work, it follows that without knowledge of the context and aims of persuasion, the method of persuasion cannot be regretted or reproached as asymmetrical and falling short of the ideal of the symmetrical communicative rationality.

The Habermasian distinction between strategic and communicative rationality was introduced in order to arrive at material guidelines for persuasion specifically, as distinct from communication generally. However, the result is a claim that persuasion should become as much a form of convincing as possible; in other words, persuasion is in principle a bad phenomenon and should be avoided as much as possible. At the same time, it remains unclear why persuasion is that bad. It "*seems* (my emphasis) to fall exactly in the middle between" strategic and communicative rationality, but Spahn doesn't explain why he thinks so and any positive characterization of a principle or basic mechanism of persuasion is lacking. This is of course understandable, because it is clear from the previous chapter that

[38] This guideline is an important improvement on the golden rule of persuasion proposed by B&N, for it centers on (factual or counterfactual) consent on the part of the recipient of persuasion and not, as B&N do, on the hypothetical consent of the persuader.

such positive characterization is an enterprise on its own. Still, without such positive characterization it is in my opinion impossible to make such far reaching claims about (technological) persuasion.

From the perspective taken above (chapter two), persuasion is not prima facie inferior and problematic, so, the real question is not how to limit persuasion but to ask what are acceptable forms of persuasion, in what situations, and for what purposes. In the next section I will take up this task (in which the four validity claims will prove useful) and develop a framework for the assessments of PTs, building on all four proposals discussed so far.

3.2 A framework for an ethics of PT.

3.2.1 The framework

Based on reworking elements of the frameworks discussed in the previous paragraph, I propose the framework represented in figure 2 below.

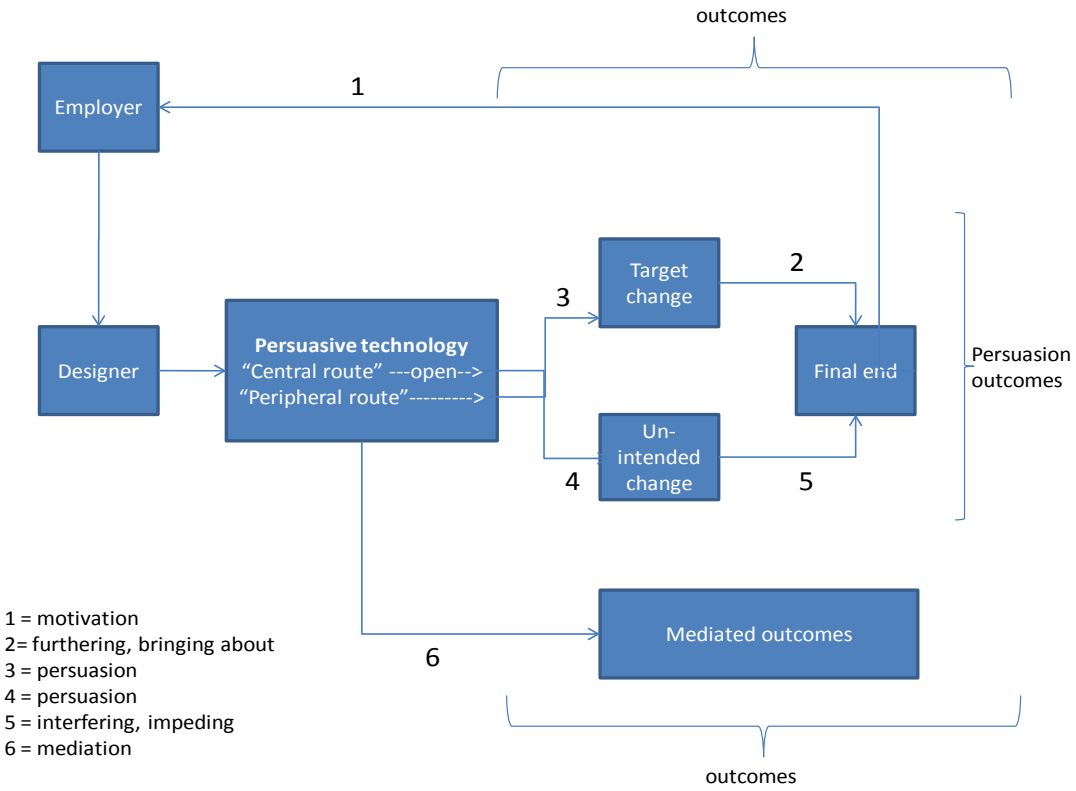

Figure 2. Framework for the ethics of PTs.

All the relations and boxes depicted in this schema need attention in ethical reflection. I will illustrate and explain the framework with the help of the EcoFeedback case given in the introduction; the { } squares used refer to the schema. The most natural way the design process starts is to note a certain problem in terms of an end or value that needs support. For example, the way the world currently uses its oil stores is unsustainable for ecological, economical, and political reasons. This insight is broadly shared now, and as a consequence many different parties (citizen, governments, industry) are motivated to make our energy consumption more sustainable, thus sustainable energy serves as a {final end} for these parties, which motivates ({1}) them i.e. the {employers} to search for means to reach that aim. It is generally acknowledged that changed patters of consumer behavior, e.g. eco driving ({target change}) in case of car drivers, is one of the solutions to the energy problem.

This is where {PT with persuasive methods} comes into the picture. The {employers} give commission to the {designers} to develop a {PT with persuasive methods} that is likely to persuade ({3}) car drivers to change their driving behavior in such ways ({ target change}) that less fuel is used ({final end}). This presupposes that there exists a clear relation ({2}) between {target change} and {final end}; this relation ({2}) needs to be firmly established for the whole enterprise of designing the relevant PT to make sense. The {designers} face a multifaceted task: they have to make a justified assessment that the {persuasive method they built into the PT}, which in case of EcoFeedback can include evaluative social feedback, will indeed lead to the {target change}, i.e. eco-driving. They also have to assure that a failure to persuade ({3}) the driver doesn't lead to very undesirable outcomes above to the failure to contribute to sustainable energy ({final end}). In this case of EcoFeedback these undesirable outcomes that should be excluded beforehand could include distraction of the driver caused by additional information and features in the dashboard, thus leading to ({6}) a decrease in safety. Furthermore in other types of PT in which functionality *depends on successful persuasion* ({3}), bad consequences may also obtain. What distinguishes relations {3} and {4} from {6}, is that the first two relations depends on the *persuasive message* communicated by the persuasive methods, whereas {6} does not, though it may affect the persuasion, e.g. by causing distraction. Relation {6} captures also the insight that every technological artifact in use influences the way we perceive the world and act in it; {6} is analogous to Verbeeks concept of mediation.

Another possibility is that next to successful or unsuccessful persuasion ({3}) for the {target change}, some {unintended changes} in driver behavior and attitudes result from the process of persuasion ({4}). For example, drivers may, even outside their awareness, more

often use their car instead of travelling by public transport, cycling, or walking[39]. This would lead to ({5}) less or even no gain in terms of sustainable energy and likely to effects on other value's as well, e.g. safety, and health.

The framework I propose is slightly different from and slightly more developed than the ones discussed above in the following ways. Firstly, as was already noted in section 3.1.1, my framework does disambiguate between outcomes in terms of changed behavior and attitudes on the one hand and the consequences of these changes for some ends or values on the other hand. By the authors discussed above, both are implicitly distinguished, but at the same time confusion can arise because they are both regularly referred to as "outcomes of persuasion".

Secondly, the role that "designer intentions/motivations" play in the other framework is in the present framework replaced by "final end". This is because the moral acceptability of a PT, its methods, and the target change is not guaranteed by good designer intentions, but needs a more objective, real world relation between a final end and the methods and target change justified by and productive to that final end. To see, look again at B&Ns' example of the three designers who built a persuasive technology to persuade some stranger to eat more fruit and vegetable The designer intentions are relevant for judging their moral virtue as a person and professional, but the moral acceptability of the persuasive technology depends on which of the three designer motivations are final ends that *are* reasonably predicted to be served by target changes that in their turn *are* reasonably predicted to be reached by the applied persuasive methods. It could be that the PT developed by the designer who secretly hopes that the stranger gets sick in his stomach by eating more fruit, in reality, in all reasonable cases and usages,only serves to reach the aim of improving his health. The designer is responsible and reproachable for his evil intentions, but the PT itself, provided its good effects were indeed reasonably predictable, is ethically acceptable.

Thirdly, the framework includes in addition to designers also employers, because usually, employers issue the design of a PT, or in any case, designers envisage certain potential employers of the PTs they develop. Designers and employers have different responsibilities as will be discussed below.

Finally, relation {6} points to the fact that technologies can have effects or outcomes that are not brought about by intended or unintended persuasion. This relation is more or less parallel to what Verbeek calls the 'otherwise mediating role of the technology'.[40]

[39] Compare Verbeek & Slob [2006] for an analogous example concerning energy saving light bulbs.
[40] In addition to persuasion and otherwise mediating the relations between humans and reality, artifacts have still other ethically relevant features, such as resources used and waste produced by its manufacturing, safety, and still others. These further aspects are not represented in my framework, because they apply to each artifact and are not specific for PTs.

3.2.2. Employers, designers, methods, target change, recipient, and final end and their ethical interrelatedness.

Now that the parts and relations of the framework are set out, it is time to elaborate on their ethical relevance (focused on methods). Because the process of designing is usually initiated by an employer with a certain final end, these will be the starting point of reflection.

Employers and their final ends, *and designers*.

A first demand on the design of PTs is that the final end needs to be morally acceptable. This is not a severe demand, for if we draw the comparison with human-human persuasion we see humans pursue all kinds of ends and heavily use persuasion as means. The law forms a minimal constraint on acceptable ends and as long as ends don't conflict with the law, they are in principle permitted to pursue. Because in the pluralistic Western democracies, agreement on which aims are ethically acceptable to pursue, will often not be reached, and as a consequence, often only the legal minimum demands need to be met.

The employer can be the government, a private party, a semi-public organization, a commercial party, or even the recipient or user himself. Often the employer will commission the design of a PT, or in case of designers who envisage a certain useful PT, they will likely cooperate with potential employers of the PTs the design. Or at least this would be wise to do, because employers often embody specific knowledge regarding the recipients and the way their behavior affects the final end in question, which implies knowledge regarding the target change of the PT. This is ethically important, because, as stressed by Berdichevsky and Neuenschwander, designers have the responsibility to make informed and reliable predictions regarding the way the PT works out in the use-practice. Knowledge of the relations between potential target changes and a given final end is also needed for reasons of efficiency: which target change is most efficient (and still ethically acceptable) in serving the final end?

Another important role of designers follows from their being expert on the methods of technological persuasion. They are in the best position to assess the actual voluntariness that the persuasive methods allow, and consequently should inform ethicists and cooperate in ethical reflection on the PT they are designing.

Target change and recipients

As with the final ends, the target change must be itself ethically acceptable; this is so even if the end is not only ethically acceptable but praiseworthy, because as a matter of fact, means are not automatically justified by the ends.

Another issue regarding recipients has to do with distributive justice. Take again the example of the value of sustainable energy. A government has to decide on how to direct its efforts to change the behavior of its citizen in sustainable directions. One possibility, eco-driving, implies a loss for those drivers who value driver-autonomy and the fun of driving vigorously from time to time. If the government would only direct its efforts at eco-driving, this would place the burden of sustainability on the group of car-drivers, and not on the other citizens, and within the group of car-drivers, on the sub-group that is persuaded. This group of car-drivers who are successfully persuaded start by definition voluntary with eco-driving, and therefore judge that they have good reasons to comply. Still, it might be interpreted as a distributive injustice that this group accepts the burden of a problem that equally concerns all.

Methods of technological persuasion

The ELM perspective on persuasion (chapter two) points to the ethical relevance of the two routes to persuasion, the central and the peripheral, and to their mutual relation. As was argued, in order to qualify as persuasion, the central route to persuasion must be open: it must be possible for users to assess their reasons for compliance with the target change by engaging in high amounts of issue-relevant thinking. This possibility must be real: users who are both motivated and able will engage in effortful elaboration. However, not all users will be motivated and/or able to engage in high amounts of thinking. Furthermore, the literature on persuasion (as cited in chapter two) suggests ways (by acting on motivation and ability) to decrease the elaboration likelihood, and thus to encourage the peripheral route. This brings up the question whether the central and the peripheral route always lead to the same persuasive outcome. Is it possible to design a PT that is highly successful in so arranging conditions that the elaboration likelihood is low and the peripheral route thus taken will lead to the target change, while this change would probably not have been brought about via the central route?

In that case, it can be argued that the designer's intent is manipulative, or comes close to being so: conditions are arranged such that the recipient is persuaded in ways in which he is largely unaware. From the truthfulness condition of discourse ethics, a guideline for the design of PT can be developed which rules out this kind of effectively manipulative PT. Truthful PT does perhaps requires that the central route and the peripheral route lead to the same outcome of persuasion. In that case, it does not matter which psychological processes do the *actual work* of change, because in either case the resulting change is the same and one which the recipient has reason for. This is however a quite stringent requirement, for often designers don’t have such control on what possibly can happen in the mind of the recipient for both routes that are possible to persuasion.

A more promising application of the truthfulness condition will be to the designers. In discourse ethics, truthfulness refers to the honesty of the speaker [Spahn, 2011] and if truthfulness is interpreted as a qualification of the designer, it rules out skillful attempts to design PTs that come close to manipulative technologies. Because truthfulness is a *prima facie* valid demand on one another in general, designers of PT are included.

One consideration perhaps shows that this fear of manipulation is somewhat exaggerated. When the persuasive issue is personally relevant, the ELM predicts that the recipient's motivation to elaborate increases. In other words, manipulation by PT will probably only succeed in matters that are not very relevant to the users, so, they will not be severely disadvantaged or ill-treated.[41] But, perhaps the mere fact of being (or knowing to be) manipulated, even absent further negative consequences, has moral weight.

'Comprehensibility' is another implicit validity claim that according to discourse ethics is made in every act of communication. In order to enable central route change, the persuasive message needs to be comprehensible. Peripheral route change typically proceeds through processing some heuristic cues. Recall the example of 'ambient persuasion trough light' (section 2.4). The soft background light serves as a simple feedback cue, most likely processed via the peripheral route. This type of ambient PT has proven effective under experimental conditions [Ham & Midden, 2010.] , which seems to imply that the cultural meanings of the colors green and red are causally effective under conditions of low thinking. Thus, comprehension in the sense of the activity of consciously perceiving and understanding the colors, and deliberately acting on such understanding is not necessary. But the fact that the change does work through the cultural meanings of the colors forms justification to regard this type of ambient persuasion as comprehensible, no matter which route to persuasion is the actual one. Questions left for future research are whether comprehensibility can differ for both routes and what the ethical relevance of such difference would be.

[41] This is perhaps a proper place to discuss B&N's 'disclosure principle': "The creators of a persuasive technology should disclose their motivations, methods, and intended outcomes, except when such disclosure would significantly undermine an otherwise ethical goal" [B&N, p 57]. Some remarks can be made. First, from the perspective of the ELM, it follows that the actual processes of persuasion can be various; still, the designer will intentionally add certain peripheral cues to the persuasive message, which they could disclose. Second, as already argued, designer motivations are less relevant to users than the actual effect of the target change on what I call the final end. Third, without specification of the 'ethical goal', it is impossible to know whether this goal justifies an exception to the disclosure principle (assumed that it is valid anyway): again, the end does not automatically justify the means (methods). Fourth, compared to human-human persuasion, the disclosure principle appears to be extremely demanding: persuasive intent, methods, and motivations of human persuaders might easily be inferred, but they are generally not explicitly disclosed. It seems not immediately evident which relevant differences between technological and human-human persuasion could justify more stringent demands on technological persuasion.

Although, on the definition developed in chapter two, PT allows for voluntary change, in practice this voluntariness will be a matter of degree. Fogg has shown that PTs can be designed to control the interaction, have generally great capacities, and have potentially a wide range of access. These three sources of asymmetry between user or recipient and PT will in practice determine the actual voluntariness for change. (The voluntariness can move on a scale ranging from 'complete' to 'close to coercion or manipulation').

The ethical acceptability of the amount of asymmetry between PT and users, and the degree of voluntariness for change that the PT permits, will have a relation with the employer and with the final end that is served by the target change. If the end (e.g. sustainability) is democratically agreed upon as even justifying coercion, then strongly asymmetrical PT is most likely justified as well; the government is the employer in such cases. If the technological persuasion is self-imposed, such strong asymmetry is ethically acceptable as well, as long as this asymmetry does not originate from a limitation of the interaction possibilities. For such PTs, the user, who is at the same time the employer, must be able to adjust the persuasion any time to (changes in) his values and his aims.

In cases of commercial employers, the final end lacks such democratic or user support. Even PT employed by the government is, by virtue of democratic control, self-imposed in a derived sense (although often citizen will not perceive it this way). In case of PT employed by commercial parties, the final end of the PT will very often not be shared by the recipient. At least not if we identify this final end as 'commercial profit'; of course, the final end of such PT could also be construed as 'helping the consumer to make an informed choice', but this seems to be more a means to profit. In any case, my intuition is that in order to be ethically acceptable, these PTs should be far less asymmetric than the two other types of PT.

Matters here are complicated however. What if on the one hand the central route to persuasion leads to favorable attitudes to the products that an artificial intelligent commercial sales-agent tries to sell, but on the other hand, the real work of motivating the consumer to actually buy the product is done via the peripheral route? What if this peripheral route is made highly effective by virtue of the huge capacities of PT, for example by the application of 'persuasive profiling' (this is the gathering and application of knowledge on which persuasive strategies works best for each individual consumer [Kaptein and Eckles, 2010])? I leave also these questions for future research.

3.3 Persuasive technology, autonomy, harm to others, and paternalism.

In the previous section, the focus was mainly on procedural issues: aspects of ethical reflection and some of their mutual relations were briefly discussed. In this section, I want to

contribute to a more substantial evaluation of PT in general, by investigation how PT relates to the value of personal autonomy.

Persuasive technology belongs to the more general class of 'behavior-influencing technologies'. Proponents of this class of technology argue that in a society that is increasingly shaped by technology, we should delegate part of our moral decision-making to this technology. The Dutch Philosopher Hans Achterhuis argued for this stance in developing his idea of 'moralization of technology' and his plea was taken over and defended by, amongst others, Peter-Paul Verbeek, who speaks of 'materializing morality' [Verbeek, 2006, p369]. Others have vigorously attacked these proposals by arguing that they amount to an unwarranted interference with our autonomy and to technocracy instead of democracy.[42]

In this section, I will discuss the autonomy issue, but only in relation to the narrower class of PT. With the help of Feinburg's discussion of the meanings of the term 'autonomy', I will first argue that, *prima facie*, PT that falls within the definition of the previous chapter does not interfere with autonomy. Then I will discuss the principles of harm and of paternalism that are proposed as justification for limiting autonomy, as they might apply in the context of PT. I will conclude that threats to autonomy mainly arise from the possibility that PT is in fact not persuasive but manipulative or coercive instead.

Feinburg [1989/86] distinguishes 'four closely related meanings' of 'personal autonomy': i) the *capacity* to govern oneself, ii) the actual *condition* of self-government, iii) an *ideal* of character, and iv) the *right* or sovereign authority to govern oneself. These distinctions will prove helpful in order to discover how PT might bear on user autonomy. [43] To start with, one could make a very straight-forward argument for the conclusion that PT does not at all interfere with or limits our autonomy. For, both on the proposed definition that demands the 'possibility of central route change as on the standard definition, that demands 'voluntary change', PTs leave us free to deliberate, choose, and act accordingly. So, PT does not deny us the right to autonomy (iv). Nor does PT by definition have a negative impact on our actual condition of self-government (ii). PTs can have such impact, but only if users let that happen, if they do not properly use their broadly rational capacities for self-control. This possibility is, however, nothing special about PT, because people can fail to act autonomously in a myriad ways under conditions that allow for autonomy.

[42] See Brey [2006] for a discussion.
[43] Of course, developing my argument on the basis of only these definitions will give this treatment of the relation between PT and autonomy the character of a preliminary investigation. A proper discussion would give conceptualizations of each of these four meanings, show how they relate, and should make clear why autonomy is important, and, accordingly, what conditions justify its limitation. However, I have no time and space for such discussion here.

This way of viewing PT as no threat to autonomy is in line with Mill, who holds that the "own good, either physical or moral, of a man" "[…] are good reasons for remonstrating with him, or reasoning with him, or persuading him, or entreating him, but not for compelling him or visiting him with any evil in case he do otherwise" [Mill, 1985/1859, p 68; see also p163]. So, for Mill compelling but not persuading interferes with an individual's liberty and individuality (Mill's terms that are more or less equivalent to the present notion of autonomy).

Nonetheless, I see two important ways in which autonomy can be limited in the context of PT, both of which have to do with the characteristics of technological methods of persuasion. In the first place, in many cases of PT, the user cannot avoid the *attempt* of persuasion. In case of human-human persuasion, typically you can negotiate with the persuader or simply walk away, but you have to put up with, for example, your eco-feedback dashboard. In the second place, the issue of autonomy is relevant to PTs, because although by definition PT allows for autonomous change, in practice it is always an empirical question whether a purported PT *de facto* employs persuasive, and not coercive, or manipulative methods. In terms of the ELM, it is an empirical question to what extent the central route to persuasion is open. For, personal autonomy requires that I am able to deliberate on what reasons I have for change and to act upon the outcome of that deliberation.

The principles of harm, and of paternalism [Beauchamp, p 389] that justify autonomy limitation are relevant in case of PT. Mill's harm principle states that state coercion of the individual is only justified in order to prevent harm to others. For the eco-feedback dashboard example, one could argue that car-driving is damaging to others, because of the emission of fine-dust, and its contribution to the climate problem. Therefore, coercive measures to decrease the damage done to others by car-driving are justified. A legal coercion to build eco-dashboards in every car is just such measure, but by no means as limiting to driver-autonomy as could be justified based on the harm-principle. For, drivers are limited in their autonomous choice of car-interior, but are as free as before to drive the way they do. Interestingly, if coercion is justified in this context, then perhaps persuasive methods that are on the coercive or manipulative end of the spectrum, which are no longer persuasive methods proper, are probably be justified as well, although further investigation of this issue is needed.

An important class of PTs is designed for well-being, for example the already discussed weight-loss website, or devices that persuade patients to compliance with their medication schemes by giving signals if the medicin is not taken in time. [IJselstein et. al., 2006]. Insofar this PT limits or interferes with the autonomy of its users the justifying ground will be the principle of paternalism. According to Dworking [2010], "Paternalism is the interference […] with another person against their will, and defended or motivated by a claim that the

person interfered with will be better off or protected from harm"[44]. Again, PT by definition leaves the user free and does not limit his autonomous choice, but if PTs designed for well-being would be coercively present, than that presence is a form of paternalism. Such interference could be justified from a soft-paternalistic view, which holds that coercive interference is needed in order to investigate whether the person in question acts voluntary (the paradigm example is Mill's one of the man who unknowingly wants to cross a damaged bridge) [Dworkin, 2010]. Because technology plays an important role in such cases, the term 'technological or technology paternalism' does apply.[45]

Instead of being a threat to autonomy, PT can also enhance autonomy: in some cases, PT can support our capacities for self-government (see meaning (i) above) and help us to reach more closely the character-ideal related to self-government (iii). Especially the (dominant) class of PTs that intend to promote attitude-behavior consistency (see section 2.5.2 above) seems powerful support for one's autonomy *when applied to oneself*. For example, the weight loss website may both enhance one's self-discipline and provide task support. *Acting* upon existing positive attitudes toward losing weight and eating healthy surely counts as self-government.

To conclude this section, the relation of autonomy with PT is a topic worthy of further study[46].

[44] His conceptual analysis provides the following conditions. X acts paternalistically towards Y by doing (omitting) Z:

- Z (or its omission) interferes with the liberty or autonomy of Y
- X does so without the consent of Y
- X does so just because Z will improve the welfare of Y (where this includes preventing his welfare from diminishing), or in some way promote the interests, values, or good of Y.

[45] Cf. Hofman [2003] who argues that " 'Technological paternalism' expands the traditional conception of paternalism beyond intentional reduction of individual autonomy to also include altered autonomy due to epistemological and societal frameworks (such as technology)". Spiekerman and Pallas [2004] argue in a context of ubiquitous computing that autonomous technology that cannot be overruled without sacrificing functionality is a core feature of 'technology paternalism'.

[46] For example, Spiekerman and Pallas [2004, p9] introduce the notion of 'perceived paternalism'. Though it seems at first sight that a *real* limitation of your autonomy both is more important and has more weight in normative considerations, *perceived* limitations of autonomy could perhaps be relevant as well. A further issue, as already brought up, is the actual autonomy effect of what is purportedly an instance of PT.

4 Conclusion and outlook to further study

Throughout this thesis, empirical knowledge of the psychological processes that underlie persuasion proved to be of central importance to understanding the phenomenon of persuasive technology, to developing a principle and a definition of persuasion and persuasive technology, and to generating ethically relevant distinctions and questions. From the perspective of the elaboration likelihood model of persuasion, I argued for an underlying core or principle of persuasion. Persuasion, which always involves an act of communication, is characterized by the openness of the central route to persuasion. That is, persuasion leaves open the *possibility* for recipients to assess their reasons for compliance with the target change by engaging in high amounts of issue-relevant thinking. This definition enables the distinction with convincing, manipulating, and coercing, and furthermore, it conceptually unifies these instances of PT that actually are of a persuasive nature. By virtue of defining persuasion in relation to a *possible* route of change, and not in relation to a specific *actual* mechanism of change, the definition can both account for the broadness of the phenomenon of persuasion and provide conceptual unification.

Ethical questions concerning methods of technological persuasion that arise from the ELM perspective center on the two routes to persuasion. Much ethical reflection on concrete instances of PT and its methods will aim at investigating the *actual openness* of the central route, i.e. at investigating whether what is called PT is indeed PT. Methods of technological persuasion need to be assessed from the perspective of the broader framework for ethical reflection on PT, which I developed by extending and modifying four existing frameworks. The central parts of this framework are the final end, which should be served by the target change (in behavior or its mental determinants), which change is brought about by the persuasive methods. Each of these should be morally acceptable in itself, and in addition, designers and employers should also establish these relations of ‘serving’ and ‘bringing about. At the same time, the occurrence of unintended changes as result of technological persuasion, and of unintended mediation effects need to be excluded or circumvented. With regard to the way PT relates to personal autonomy, I argued that as long as PT is de facto persuasive, and if the use of PT is not required, the personal autonomy of the user is *prima facie* not threatened.

Many issues that require further study were identified already, and many can be added. These issues can be grouped into a psychology, a philosophy, an ethics, and a political philosophy of persuasive technology, though many mutual relations between these

sub domains do exist. The study of the psychology of persuasion will benefit from considering other models of persuasion, and more general literature on dualsystem theory. For example, the question regarding the mutual relation between the peripheral and the central route should also be treated from the heuristic-systemic model of persuasion, for this model treats both routes much more complementary and potentially co-occurring than the ELM does. Of great import is the rapidly growing insight in the roles of automaticity, and the unconscious.

If a PT can trigger automatic attitude activation, of which the user is not aware, does this still count as PT? This question is part of the project of further developing a philosophy of PT. The distinctions between technology that persuades, convinces, manipulates, or coerces, need further attention. The relation and distinction between communication and persuasion, in combination with the role of designer intentions, obviously need much more elaboration then was given in this thesis.

The ethics of PT will build heavily on the psychology and philosophy of PT. Should there be *some kind of* correspondence between the outcomes of central route and peripheral route change?[47] If so, why and what kind of correspondence or congruence? Or is the possibility of correspondence sufficient? What if a PT strictly or formally allows for central route change, but which is explicitly designed to make it very likely that the majority of users will follow the peripheral route, for which the PT provides cues that make certain intended outcomes of persuasion probable? Much more work needs to be done on under what conditions central rout persuasion is possible. A further question regards how our capacities for practical reasoning are affected by PTs. And, what about creating the appearance of tailoring, without actually tailoring the message to the recipient [Fogg, 2003, p 40]

Finally, a political philosophy of PT should answer the question for what societal ends PT is the appropriate means. For, other means of influencing citizen behavior are available: primary education, legal coercion, and public information campaigns. Which groups will be the target audience? And, are the burdens of changing behavior into societal desired directions justly distributed? A political philosophy of PT becomes especially urgent when the use of PT is made legally required. An important question is whether justification of coercive legislation by the same token forms justification for legally required PT.

Many questions thus ask for an answer. As will be clear from the way I presented them, answering them will benefit greatly from interdisciplinary cooperation between psychologists, designers, philosophers and ethicists.

[47] I deliberately write 'some kind of' correspondence, because I already argued that 'robust' correspondence is too strict a demand. As an illustration of this point, consider a very simple case. A message containing ten bad arguments could lead to compliance via the peripheral route, via the 'great number of arguments must be true' cue. Via the central route, the badness of the arguments will be noted, and no compliance will result.

List of cited literature

1. Ajzen, I., Fishbein, M., 2000, Attitudes and the attitude-behavior relation: Reasoned and automatic processes, *European Review of Social Psychology* (**11**) pp. 1-33

2. Baker, S., Martinson, D.L., 2001, The TARES Test: Five Principles for Ethical Persuasion, *Journal of Mass Media Ethics* (**16**), pp. 1480-175

3. Beauchamp, T.L., , 1991, *Philosophical Ethics. An Introduction to Moral Philosophy*, 2nd edition. New York:McGraw-Hill

4. Berdichevsky, D., Neuenschwander, E., 1999, Toward an Ethics of Persuasive Technology C*ommunications of the ACM* 42, 51-58

5. Brenkert, George G. 2008. *Marketing Ethics*. Malden:Blackwell, pp. 155-156

6. Brey, P., 2006, The ethics of behavior-steering technology. In *User Behavior and Technology Development*, edited by P. P. C. C. Verbeek and A. Slob. Dordrecht:Springer

7. Cialdini, 2007, *Influence: The psychology of Persuasion*. New York:Harper Collins

8. Dance, F. 1970, The concept of communication, *The Journal of Communication* (**20**), pp. 201-210

9. Davis, J., 2010, Generating Directions for Persuasive Technology Design with the Inspiration Card Workshop in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 262-273, 2010

10. Dworkin, Gerald, "Paternalism", *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/paternalism/>.

11. Feinburg, J. 1989, Autonomy, in: Christman, J. (ed). *The Inner Citadel. Essays on Individual Autonomy*, New York;Oxford:OUP. (original publication in 1986)

12. Fogg, B. J. 2003. *Persuasive Technology: Using Computers to change what we think and do*. The Morgan Kaufmann series in interactive technologies. Amsterdam; Boston: Morgan Kaufmann Publishers

13. Ham, J.R.C., Midden, C.J.H., 2010, Ambient Persuasive Technology Needs Little Cognitive Effort: The Differential Effects of Cognitive Load on Lighting Feedback versus Factual Feedback. in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 132-142, 2010

14. Hofman, B, 2003, Technological Paternalism. On how medicine has reformed ethics and how technology can refine moral theory. *Science and Engineering Ethics* (**9**), pp. 343-352

15. Kaptein, M., Eckles, D., 2010, Selecting Effective Means to Any End: Futures and Ethics of Persausion Profiling in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 82-93, 2010

16. Kruglanski, A.W., Thompson, E.P., 1999, Persuasion by a Single Route: A View From the Unimodel, *Psychological Inquiry* (**10**), pp. 83-109

17. Latour, B., 1992. Where are the Missing Masses? Sociology of a Few Mundane Artefacts. In W. Bijker and J. Law (Eds.) *Shaping Technology, Building Society: Studies in Sociotechnical Change*. Cambridge, Mass, MIT Press: 225-258

18. Letho,T, Oinas-Kukkonen, H., 2010, Persuasive Features in Six Weight Loss Websites: A Qualitative Evaluation in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 162-173, 2010

19. Mitcham, C., Schatzberg, E., 2009, Defining Technology and the Engineering Sciences, in: *Handbook of the philosophy of science. Volume 9: Philosophy of technology and engineering sciences*, edited by A. Meijers. Oxford: Elsevier

20. Merriam Webster Online Dictionary, accessed on 5/7/11 'to persuade', URL: http://www.merriam-webster.com/dictionary/persuade

21. Meschtscherjakov, A., Wilfinger, D., Sherndl, T., Tscheligi, M., 2009, Acceptance of Future Persuasive In-Car Interfaces. Towards a More Economic Driving Behavior, in: Proceedings of the First International Conference on Automotive Usre Interfaces and Interactive Vehicular Applications. New York:ACM

22. Mill, J.S., 1958/1859, *On Liberty*, Penguin Books

23. Nickel, P. *The Definition of Persuasive Technology*, work in progress.

24. Nickel, P.J., Spahn, A., *Design for Symmetry - Trust, Discourse Ethics, and Persuasive Technology*, submitted.

25. Oinas-Kukkonen, H, 2010, Behavior Change Support Systems: A Research Model and Agenda. in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 4-14, 2010

26. O'Keefe, D. 2002, 2e edition. *Persuasion. Theory & Research*. California: Sage Publications.

27. Petty, R.E., Brinõl P. Psychological Processes Underlying Persuasion: A Social Psychology Approach, *Diogenes*, 2008, (**217**) 52-67

28. Petty, R. et. al, 1999, Is there one persuasion process or morer? Lumping versus splitting in attitude change theories*, Psychological Inquiry* (**10**), 1999, 156-163

29. Petty, R.E.; Cacioppo, J.T.; Strathman, A.J.; Priester, J.R. Brock, 2005, To Think or Not to Think: Exploring Two Routes to Persuasion., in: Timothy C. (Ed); Green,

Melanie C. (Ed), *Persuasion: Psychological insights and perspectives*, 2nd ed, (pp. 81-116). Thousand Oaks, CA, US: Sage Publications

30. Roubroeks, M.A.J., Ham, J.R.C., Midden, C.J.H., 2010, The Dominant Robot: Threatening Robots Cause Psychological Reactance, Especially When They Have Incongruent Goals, in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 174-184, 2010

31. Rawls, J, 1999, *A Theory of Justice*, Massachusetts:The Belknap Press (original edition: 1971)

32. Ruijten, P, Midden, C, Ham, J., manuscript, Unconscious persuasion needs goal-striving

33. Spahn, A, forthcoming, And Lead Us Not Into Persuasion, …? Persuasive Technology and the Ethics of Communication, *Science and Engineering Ethics*.

34. Spiekerman, S., Pallas, F., 2006, Technology Paternalism - Wider implications of Ubiquitous computing, *Poiesis Prax* (**4**), pp. 6-18

35. Strack, F., Deutsch, R, Reflective and Impulsive Determinants of Social Behavior, *Personality and Social Psychology Review*, 2004 (**8**) 220-247

36. Van de Poel, I., 2009. Values in engineering design. In *Handbook of the philosophy of science. Volume 9: Philosophy of technology and engineering sciences*, edited by A. Meijers. Oxford: Elsevier

37. Verbeek, P.P.C.C., manuscript, *Persuasive Technology and Moral Responsibility: Toward an ethical framework for persuasive technologies*. Paper read at Persuasive06, Eindhoven. Available from: http://www.utwente.nl/gw/wijsb/organization/verbeek/verbeek_persuasive06.pdf (last accessed on 29-08-2011)

38. Verbeek, P.P.C.C. 2006. Materializing Morality. Design Ethics and Technological Mediation. *Science, Technology and Human Values* 31 (3):361-380

39. Verbeek, P.P.C.C., Slob, A, 2006. Technology and User Behavior: An Introduction. In *User Behavior and Technology Development*, edited by P. P. C. C. Verbeek and A. Slob. Dordrecht:Springer

40. Vossen, S., Ham, J., Midden, C., 2010, What Makes Social Feedback from a Robot Work? Disentangling the Effect of Speech, Physical Appearance and Evaluation, in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (eds.), *Persuasive Technology. 5th International Conference, Persuasive 2010, Copenhagen, Denmark, June 2010, Proceedings*, Berlin, Heidelberg and New York: Springer, 52-57, 2010

41. Waelbers, K., 2007, What Things Do: Philosophical Reflections on Technology, Agency and Design. A Book Review. *Science and Engineering Ethics* 13 (2):275-277

42. Wilson, Timothy D. 2002, *Strangers to Ourselves. Discovering the Adaptive Unconscious*, Cambridge:Belknap Press

43. IJsselstein, W., Y. De Kort, C. Midden, B. Eggen, & B. van den Hoven, 2006, Persuasive technology for human well-being: Setting the scene. in: IJsselstein, de

Kort, Midden, Eggen, van den Hoven (Eds.): *Persuasive Technology. First International Conference on Persuasive Technology for Human Well-Being* (Proceedings), Berlin, Heidelberg and New York: Springer, p. 1-5., 2006

44. http://world.honda.com/news/2008/4081120Ecological-Drive-Assist-System/ accessed on 17 august 2011