Trust, Discourse Ethics, and Persuasive Technology

Philip Nickel, Andreas Spahn

Ethics and Philosophy of Technology, TU/Eindhoven

Abstract. In this paper we analyze the role trust plays in an ethical evaluation of PT. We distinguish between trust in PT itself, and trust in those humans who design, produce and deploy it and draw on Discourse Ethics to further distinguish two types of communication embodied in PTs: *asymmetrical* and *symmetrical* communication.

Keywords. Persuasive Technology, Trust, Ethics, Discourse Ethics, Speech Act Theory

1 Introduction

PT raises important issues about trust, because 'trusting' a PT seems to go further than 'trusting' a non-persuasive technology. Users are required to put a more richly normative or even anthropomorphic trust in PTs: one trusts that the technology 'knows' which behavior is – for moral reasons – more adequate and what might help one to adopt that behavior. Such technology invites us, to some extent, to trust it in moral matters.

2 Persuasion and Discourse Ethics

It seems reasonable to classify PT as a communicative or technology, since it mirrors some aspects of human communication. Firstly PT aims at *transferring a message*. Often this message implies an action-guiding imperative. A blinking light in a car might be designed to express something to the effect of: 'Put on your seatbelt!' The second aspect that PT shares with human communication is that it is intended to result in a change of behavioral dispositions and corresponding attitudes. Of course not all human communication is directed at this aim. However, ordinary 'ethical communication' is intended to convince others of the shared value of certain desired behaviors. Therefore a Discourse Ethics (DE) framework is an especially promising means of evaluating the ethics of PT [1].

2.1 Symmetrical and Asymmetrical Communication.

DE distinguishes two types of rationality linked to two types of intersubjective relations. C*ommunicative rationality* is a symmetrical relation between humans; *strategic rationality* presupposes (or establishes) an asymmetrical relation [2]. Both types of relations may result in a change of behavior – but whereas communicative rationality aims to *convince* the other that a certain behavior is desirable; strategic rationality aims to *manipulate* him so that

behavior change is a foreseeable effect. Communicative rationality asks for the rational assent of others concerning what 'ends' or aims we should have, while strategic rationality is about finding the most effective strategy to reach a given outcome or end. When communication is merely strategic, it does not involve communicative rationality, for in this case communication is no different than any other means of achieving behavior change, such as coercive threats or pharmaceutical intervention.

Symmetrical communication allows each party to the communicative act to evaluate and respond to the relevant reasons for action or belief put forward in that act. This response need not be overt; it can consist of a private evaluation of the relevant reasons and the appropriate response. But the possibility of an overt response is an important indication that symmetrical communication is present. DE tries to establish moral rules that indicate whether or not a given discourse may or may not count as an exercise of communicative rationality [3-4]. According to DE, ethically permissible communication about norms and values has to adhere to communicative rationality, and for that reason it must be symmetrical. This implies that only sound arguments should be part of this communication and all relations of hierarchy and authority are inappropriate so long as they prevent rational communicative responses and criticism. The purpose and outcome of the discourse are open in a strong sense, because any party to the communication could be convinced by the other parties to change their behavior or their moral beliefs.

Asymmetrical communication excludes the possibility of reply and criticism. In cases where the purpose is merely to provide information, this is ethically unproblematic: directional signs, for example, communicate asymmetrically but are ethically unproblematic. When used for the express purpose of behavioral change, however, asymmetrical communication is a form of strategic rationality, which is not about finding or agreeing upon shared values based on shared premises, but rather uses the other – as Kant would say – as an instrument for one's own practical goals. Speech and meaning are not addressed to the other person's independent practical reasoning, and so the other person's autonomous practical judgments are thus not respected as intrinsically important. I try to get the other to act so that his behavior matches my aims, but it is purely incidental whether he shares or agrees with my intentions. Strategic rationality thus has its clearest form in pure manipulation and the exercise of power. Strategic rationality can result in behavior change, but by ignoring the independent judgment of those it seeks to change, according to DE it is a morally unacceptable way of reaching this aim.

'Persuasion' falls in between symmetrical and asymmetrical communication, since it does not primarily or exclusively appeal to reason or arguments, nor does it use purely manipulative techniques of behavior change. Indeed the question, under which conditions persuasion can be distinguished from manipulation or propaganda, is an important topic in the ethical literature on PTs [5]. Persuasion seems to share with strategic rationality the emphasis on finding an 'efficient' mechanism of behavior change, and it seems to share with communicative rationality that it tries to do so for a moral reason – assuming that the user either shares the value in question, or would do so if she were fully rational. *Designers* thus seem to have two very different tasks when creating PT: the search for an efficient mechanism on the one hand, and making sure that the PT contributes to a moral value that the user does or would in principle agree upon, on the other. Let us call the second task the *sincerity requirement*. This requirement goes beyond the usual ethical requirements placed on designers. Looking at PTs from the user's perspective, one can see why.

3 Trust and Designing for symmetry

Trust in technology consists of two elements: a judgment that the technology is sufficiently likely to perform a certain way to be worth relying on; and a normative expectation that one is *entitled to* a certain level of performance from the technology. A reliability judgment alone is insufficient for trust because, in a failure of trust, one attributes the failure *to the technology*, not to oneself or to bad luck. A normative expectation must be supposed in order to explain this distinctive feature of trust [6].

Trust in persons also consists in two parallel components: a judgment that the person is worth relying on in a certain domain, and a set of normative expectations that she will behave a certain way in that domain. Here the normative expectations are more richly evaluative, including ethical or moral elements. One supposes that the trusted person will take one's own interests into account [7], that he or she shares your moral values [8], or that she has a moral obligation to behave a certain way [9].

PT needs to do more than merely function reliably (i.e., performing to specification under foreseeable conditions) in order to be trustworthy. It must actually match the values of the user or target individual in order to be trustworthy. In this way the trust that a user places in PT is more like trust in persons than trust in technology. As part of her trust, the user will have certain moral expectations of the technology.

On the other hand, PTs are often not designed so as to respond and adapt to the specific moral expectations of users. They do not have a capacity for symmetrical communication, because they do not incorporate the possibility of hearing, responding and adapting to the moral values of the user. This means that whatever approximation of the user's values is incorporated into the PT is often determined by the designer, manufacturer and/or deployer *a priori*, in an attempt to meet the sincerity requirement.

There are three disadvantages to this method of meeting the sincerity requirement. First of all, it places a greater moral burden on the designer, to be able to anticipate the values of the user and how the user would prefer to implement those values. If the PT itself were able to adjust to the values of the user, then the designer would not have to foresee every instance in which the value is implemented, in order fully to meet the sincerity requirement. But if the PT cannot do this, then the designer must fully anticipate the implementation of the value in every situation in order fully to meet the sincerity requirement, and to be trustworthy by the lights of the user. Practically speaking, this will be very difficult.

Secondly, the *a priori* method of incorporating moral values in PT only partially approximates the value of communicative rationality. Communicative rationality requires, not just that the speaker communicates things that the hearer can rationally accept, but also that there is the possibility of a reverse channel of communication, and the possibility that both parties might adjust their position mutually in accordance with considerations put forward during the conversation. One reason why this is important is that it is respectful to the independent judgment of both speaker and listener. A second reason is that there are often questions about how different values should be balanced in particular contexts. Thirdly, the *a priori* method provides no means for the user to test the trustworthiness of the particular PT. Although it is important for the user to have prior trust in the PT, acquired through contextual information sources or through the reputation of the designer or deployer of the PT, it is also important that the user has an opportunity to gain first-hand experience of the adequacy and fit of the PTs implementation of a given value. In the case

of ordinary artifacts, this empirical evidence of trustworthiness is acquired by engaging in normal use of the artifact, and satisfying oneself that it performs its function. In the case of interpersonal trust, empirical evidence of trustworthiness is acquired in a somewhat different way. It is partly acquired by occasionally making one's expectations and values known to the trusted person, so that she can demonstrate that she is capable of adapting her behavior to those expectations and so, evidently, has the interests of the trusting person at heart.

Therefore, we advocate making PTs more flexible and symmetrical in their design, allowing them to react and adapt to the preferences and values of the user while nonetheless striving to achieve the desired social value. This will relieve designers of the need to make *a priori* judgments of how ethical values should be implemented and balanced against each other and against other personal values.

4 Conclusion

Most PTs nowadays are designed closer to the concept of asymmetrical communication and do not meet the requirements of symmetrical communication. Very often the sincerity condition is not taken into consideration, as it is simply assumed that the user will be sharing the values that the PT is designed to promote. For the perspective of the designer it means that he has to anticipate the values the users will be willing to accept, while the user has to trust (as it were blindly) that the PT is promoting morally sound values. We have argued to place more emphasize on the sincerity condition: making sure that PT contributes to a moral value that the user does or would in principle agree upon. In principle there are two ways to adhere to it: one might be to involve the user more in the design of PT as approaches of participatory technology development have suggested [10]. But the emphasis is here to make the design situation more symmetric (between user's and designer), whereas we suggest to make the PT more adaptive to the user's needs, values and potential worries – thus not changing the design conditions, but rather the designed object itself.

5 References

1. Spahn, A. PT and the Ethics of Communication, *Science and Engineering Ethics*, forthcoming.
2. Habermas, J. *The theory of communicative Action,* Boston: Beacon Press 1984.
3. Apel, K.-O.: *Transformation der Philosophie.* Frankfurt a. M.: Suhrkamp 1973.
4. Kuhlmann, W. *Reflexive Letztbegründung.* Freiburg/München: Alber, 1985.
5. Berdichevsky, D., Neuenschwander, E. Toward an ethics of persuasive technology. C*ommunications of the ACM* 42, 51-58, 1999
6. Nickel, P. Trust in technological systems. de Vries, et al: (eds.), *Norms and the artificial: moral and non-moral norms in technology*. Springer, forthcoming
7. Hardin, R. *Trust.* Polity, 2006.
8. McLeod, C. *Self-trust and reproductive autonomy*. MIT Press, 2002.
9. Nickel, P. Trust and obligation-ascription. *Ethical theory and moral practice* 10, 309-319, 2007.
10. Davis, J. Gernerating Directions for Persuasive Technology Design with Inspiration Card Workshops, T. Ploug, et al (eds.), *Persuasive Technology.* Berlin et al: Springer, 262-274, 2010.