On Deception and Trust in Artificial Agent Development

Frances Grodzinsky, Sacred Heart University
grodzinskyf@sacredheart.edu
Keith Miller, University of Illinois Springfield
miller.keith@uis.edu
Marty J. Wolf, Bemidji State University
mjwolf@bemidjistate.edu

Abstract

This paper begins by exploring two fundamental questions: What is deception? When is it permissible to be deceptive? We explore these questions in the context of trust and the development of artificial agents (AA). Are developers using deception to gain our trust? Is trust generated through technological "enchantment" warranted? Finally, we analyze the role of trust in situations involving masquerading machines, and how deception that involves both humans and machines impacts software development in general.

1 Introduction

In her book *Alone Together*, MIT anthropologist Sherry Turkle (2011:90) writes:

> In *The Republic*, Plato says, "Everything that deceives may be said to enchant." The sentiment also works when put the other way around…That which enchants, deceives.

We are enchanted by Siri, the personal assistant on the iPhone. Robot pets and babies delivered to nursing homes thrill the patients, who treat the machines like "real" pets and babies. The deception is there: Siri and the babies are not human, and the pets are machines, not animals. Yet, even when we are consciously aware of these distinctions, the power of these deceptions causes us to interact with these devices as *if* they were people or animals (Grodzinsky et al., 2009). In the foreseeable future, many more of us are likely to be deceived by the next generation of silicon-based devices that masquerade even more effectively as carbon-based life forms.

In an earlier paper, we explored ethical concerns of deception in relationship with artificial agents (AAs). We argued that some deceptions are benevolent in that they are "helpful for the users negotiating the often complex world of technology" and that other deceptions can cause great harm (Grodzinsky et al., 2013). We concluded that the use of deception by software developers requires ethical justification, and that the ramifications of deceptions are likely to become more complex, both legally and ethically, as AAs become increasingly human-like.

In this paper we explore the relationship between trust and ethical concerns about deception in the development of AAs. We will argue that deception is a double-edged sword as it can be used both to enhance the trust relationship as well as degrade it. We are particularly concerned about machines "masquerading"--leading another agent to believe or behave as if the machine were human. There are numerous factors that influence trust, including benevolence, honesty, competence, predictability

(McKnight and Chervany, 1996), and identity, reliability, and transparency (Grodzinsky, et al., 2011). Masquerading is a particular type of deception that raises questions about trust.

Both deception and trust are widely discussed in scholarly literature. We begin this paper by exploring some of that literature, and we adopt notions of deception and trust that we find applicable to the development of AAs. We explore the following questions: Are developers using deception to gain our trust? Is the trust generated through technological "enchantment" warranted? Next, we investigate more complex questions of how trust relationships that involve deceptive AAs differ from trust relationships that only involve deceptive humans. Finally, we analyze the role of trust in situations involving masquerading machines.

2 Important Definitions and Notation

Levels of Abstraction (LoA) as developed by Luciano Floridi are an important tool for our analysis (see (Floridi, 2008) and more recently (Floridi, 2013)). Software developers have a huge responsibility in deciding if, when, and how to use deception; they are responsible for building the trust relationships that exist because of AAs. We adopt notation introduced in (Grodzinsky et al, 2013) to describe the sets of observables available to the developers (LoAD) and the set of observables available to the users (LoAU). These sets of observables clearly intersect, but one does not contain the other since developers do not typically have access to the entire context in which the user is running the software, and users do not typically have access to the technical details of the AA. We also use LoAS to describe the set of observables available on a societal level (Grodzinsky et al, 2011).

2.1 Deception

The kind of deception of interest to us in this paper focuses primarily on a "violation of principle" (Solomon, 2009:26). This implies an intentional, successful deception by developers and a misapprehension by people other than the developers. This notion is consistent with much of the literature on deception. Michael P. Lynch requires deception to be a misleading that is "*willful or non-accidental*. So, X deceives Y with regard to f only if X willfully causes Y to fail to believe what is true with regard to f" (2009:190-191). Similarly, Thomas Carson agrees that deception requires some kind of intention to cause others to have false beliefs. "A person S deceives another person S1 if, and only if, S intentionally causes S1 to believe x, where x is false and S does not believe that x is true" (2009:178 – 179). Carson says that deception connotes success, i.e., a deception is something that is believed.

If we accept the premise that deception connotes success, and if we limit ourselves to examples in which the deception is intentionally built into software artifacts, is this then always a bad thing for trust? "Sometimes other things matter more than truth. Thus, more of us would be willing to be deceived, or to deceive ourselves, if we thought that more good than bad would come of it overall, or that the matter was so trivial that the point was essentially moot" (Lynch, 2009:198). Robert Solomon's view of deception includes not only the particular deception, but also the context of the deception, the aims, intentions and character of the prevaricator (2009:26). If the software developer is not intending to do harm, but trying to establish a meaningful trust relationship for the end users of his/her product at LoAU or with the larger society at LoAS, it might be assumed that this is an instance of a benevolent deception. Yet, it still raises the question of how a developer plays out the virtue of honesty in trust relationships when developing sophisticated machines that perpetuate deception.

2.2 Trust

Like deception, much has been written about trust and all of its variants, including e-trust (see (Taddeo, 2009) and (Grodzinsky et al., 2011) for example) and social trust (see work by Falone at http://t3.istc.cnr.it, for example). We adopt Taddeo's (2009) definition of trust:

Trust is a relation between *a* (the *trustor*) and *b* (the *trustee*). NOTE: ***a*** and ***b*** can be human or artificial. A relation

(certainly in the mathematical sense, but also in the sociological sense) can involve both.

Even in the case that the relationship is between a developer and a user of an AA, trust can be bi-directional (though it isn't *always* so). Users typically trust developers to develop AAs that are predictable, reliable, and transparent. Developers typically trust users to deploy AAs in ways and environments for which they were designed. Yet, developers can deceive because they understand and control the AA and the interface it shares with the user, in ways that non-developers typically do not fully understand. This increased knowledge about the AA gives developers control and power over users, and this is significant in understanding the moral situation with respect to the trust relationship between developers and users. One question with ethical significance that a developer should consider carefully is whether the use of deception is likely to cause a user to make unwise choices because of wrongly believing that something is true.

3 Deception and Trust

Deceptions raise questions about trust. Do all deceptive acts connect to trust in the same way? In "Deception and Trust," Alan Strudler maintains that "not all manipulation in deception involves a breach of trust, and that deception that involves a breach of trust may involve a wrong that is distinguishable from that which occurs in other deceptions" (2009:139). If the AA developer is using deception in a helpful way for users, and if the user trusts the software artifact, can the deception be benevolent? If the user performs certain actions based on the trust he/she has in the artifact, and if that trust is misplaced (i.e., the developer is manipulating the end-user and does not have the user's best interests at heart), then there is a violation of trust. On the other hand, the developer can manipulate the user to do something that does not involve a breach of this trust simply by making the process opaque (2009:142). Strudler maintains that there is a morally important difference between deception that occurs through breach of trust and deception that occurs through a manipulation that is benign. (2009:143). However, Strudler asserts that deception involving a breach of trust is always wrong (2009:151).

Trust is breached when trust is established between the trustor and trustee and then violated by deliberate manipulation. But not all deceptive manipulations are breaches of trust. For example, a person might believe that her robot baby will stop crying when she picks it up. She might trust this to be true based on previous experiences with the artifact. The perception that *she* is making a difference to this AA exists in her mind, even though it is the AA's programming that causes the change. She becomes more attached to the baby believing that it has human responses to her actions. This is a deception, a manipulation, by the developer who designed a robot that induces a misapprehension by the user, but does not constitute a breach of trust. There is no established trust relationship between the developer and this user. Other manipulations that cause deceptions might include a machine designed to help calm demented patients who might cause injury to human health providers. By disguising a machine as a person or a pet, patients more readily accept this artifact than if it were more obviously a machine.

This example helps identify the main categories for analysis in the rest of the paper. There is the notion of the trust relationship between the developer and the user. This relationship is mediated

through the AA. The other trust relationship is between the AA and the user of the AA. In the case that the AA is a robot, this relationship is physically mediated. When the AA is some sort of agent accessible via the web, the relationship is electronically mediated (see (Grodzinsky et al., 2011) for further elaboration on other possible trust relationships). Each of these possible relationships can be complicated by the use of deception by the AA developers.

Assume that a machine masquerade—a deception implemented by a developer—occurs. Assume further that the consequences are at worst benign and at best (on balance) good. Is there something fundamental about deceiving people in this way that damages trust, despite otherwise good results?

In (Grodzinsky et al., 2013) we argued that masquerades do not necessarily poison the ethical well to the extent that ALL masquerades are to be condemned as ethically unacceptable. We found "that non-objectionable machine deceptions *are* exceptions and that the rule against such deceptions holds unless a valid argument overrides the rule in a specific situation." Such an argument must, among other things, demonstrate that trust has not been breached. Often this can be achieved by showing how the deception enhances attributes of trust, such as benevolence, honesty, predictability and transparency.

As we shall see in the next section, the consequences of a deception are often mixed and come from mixed motivations. Sometimes we have to examine the intentions of the developer. Sometimes we have to focus on the consequences of the deception.

4 Using deception in developing artificial agents

For most AAs, developers are interested in having users trust the AAs to do their jobs. That being the case, developers ought to carefully consider the impact of incorporating deception. Manipulation that was implemented to make an AA easier to use would not automatically constitute a wrong nor a breach of trust. However, there are concerns that arise whenever humans or AAs are deceived.

4.1 Human-Robot-Human Deception and Trust

One of the concerns regarding masquerading machines is that their inherent deception may have been included not because it was functionally required, but only because it was technically possible. Particular deceptions may be developed because people working at LoAD find it challenging and interesting to mimic human appearance and behavior.

When designing a robot, for example, the developer might be satisfied with the functionality of the robot (which is not currently masquerading), but may decide that it would be more entertaining to users if the robot had a human or animal face. For the earlier model of the robot, its non-humanness is always observable to users at LoAU. When a human-like interface is added to the later model, the developer is incorporating a deception into the robot. The developer (at LoAD) is aware of the true nature of the machine, but when the masquerade is added, this true nature is hidden, or at least obscured, from users at LoAU.

The decision to add a masquerade affects the trust relationship in at least two ways. First, there is the question of whether the user is more likely to establish some sort of trust in the AA because of the human or animal-like appearance. This is a question of the developer providing transparency. A second impact has to do with the identity of the AA. Since it is conceivable to mass produce these robots, it is possible for two or more of them to look the same, but behave differently, clearly having an impact on trust relationships.

Even people who are initially aware of a machine's non-humanness may be "enchanted" after repeated exposure. As a person becomes accustomed to the AA's behavior—it regularly answers

questions reliably, and it looks like us or something that is familiar to us—the person interacting with the AA may change her expectations of the AA. An initial skepticism about the machine may be replaced with assumptions formerly reserved for people. For example, a masquerading machine may be regarded as trustworthy at least in part because of its human-like appearance. But the human-like appearance should probably be irrelevant to questions regarding trustworthiness. Changes in behavior that might be noticed if the machine were more obviously a machine may go unnoticed if a machine is masquerading. That is, humans may give an unwarranted "benefit of the doubt" to a masquerading machine. If a person reaches a state where questioning the veracity of the information given by the AA is no longer a natural part of her interaction with the AA (i.e., a state of trust is established), it is easy for unintended consequences to occur as the user starts to make decisions influenced by a trust, based in part, on the masquerade. Thus, a robot practicing deception, through consistent high-level performance and familiar interfaces, can, in the mind of a user, move to a state in which the non-humanness of a robot is not an observable. This may be true even for a robot that regularly reminds the user that it is not a human. At some point, the reminders may not register.

Research by Sherry Turkle on robot babies and pets given to the elderly suggests that developing human-like trust relationships with robots is not farfetched. At the end of an experimental interaction period, she was unable to remove these robots as the users had become so attached to them. One user said, "She listens to me" (2011). Several recent movies include the theme of robots being accepted as "persons" by humans (for example, *I Robot* in 2004, *Enthiran* in 2010, and *Robot and Frank* in 2012*).*

Perhaps young people who are more exposed to technology will be less easily deceived or enchanted by machines. But Turkle reported the case of Callie, a ten year old, who took care of a robot baby for three weeks: "…loving the robot makes her feel more loved. She knows the robot is mechanical but has little concern for its (lack of) biology…she sees the robot as capable of complex and mixed emotions. Callie said, "When My Real Baby says, 'I love you,'…I think she really does" (2011:77).

The notion of AA identity can impact trust relationships in another way. Consider the developer who had good intentions and established an appropriate argument for using deception in an AA. Such a developer may have unintentionally created a viable mechanism for others with less commendable intentions to replicate the AA. After the original AA has established trust, would it be a breach of trust if someone accepts the robot based on its prior actions (in the original form), and then is harmed by a malevolent copy? We think that the second use, the intentionally harming use, is clearly a morally unacceptable act. However, we also contend that the original creation of a masquerading machine may have facilitated the subsequent unethical act. If the original masquerade was not essential, then it might have been better (more ethical) if the original machine had not been designed to implement the (unessential) masquerade.

Another example is the early development of e-mail systems. The unexamined assumption among early developers of the Internet and its precursors was that people using the Internet could be trusted. Today we see the impact of such a decision: significant amounts of network traffic and computer processing are dedicated to the eradication of spam from our electronic in-boxes, and phishing scams disrupt many lives. Significant care is in order when the potential impact enters into the human realm. The cautionary tale of email suggests that careful analysis under the assumption of AAs operating in an environment where untrustworthy humans and AAs are present is required to lead to the development of non-trivial counter-measures in order to prevent extensive harm from deceptions involving sophisticated AAs.

An obvious trust mistake that developers can make is that they automatically trust users to use the artifact in the way that the developers intended. This trust can be misplaced in at least two ways: users with nefarious intentions may intentionally misuse the artifact. And other, more benevolent users many not use AAs in a competent or predictable way. If a developer blindly trusts users, the developer may release an AA (particularly a masquerading AA) that will prove to be harmful due to misuse.

4.2 Humans Deceiving AAs and Trust

There are active philosophical objections to the very notion of machines trusting humans. Despite these objections, it seems reasonable to consider a world in which AAs can effectively pass the Turing test for long periods of time, behaving in ways that at least appear similar to humans in trust relationships. For our purposes, we will use the term “trust*” to indicate the machine analog to human trust. People might choose to deceive these human-like AAs. How does this impact the trust* an AA might have in individual humans?

Consider an autonomous robot acting as a guard at a dangerous area. A deception could be done involving the machine when someone seeking unauthorized access fools the robot and enters the area under false pretenses. While the robot does not have “intelligence” identical to human intelligence, in this scenario it has functional information processing capabilities that are similar to human intelligence. Depending on the level of sophistication of the robot, people may use strategies previously used to deceive people in order to deceive this machine. It seems both reasonable and accurate to use at least some of the terminology we have developed to describe human-to-human deception in order to describe human-to-machine deceptions, even though there are important distinctions between sophisticated machines and humans.

One of these distinctions is that machines may someday be better at detecting deception than most humans are. Researchers are working on software to detect micro-expressions that appear in human faces when they lie (Wilson 2011). While the researchers point out that detecting a lie is different than determining the truth, and lying is a more overt activity than deception, we consider it significant that software may someday detect deception. If AAs become likely to detect human deception, and if humans are less likely to detect AA deceptions, this will change the power balance between AAs and humans.

Should machines become capable of detecting human deception (benevolent or not), a question that developers will face is what should an AA be programmed to do when it detects a deceptive person? As machines become more sophisticated about deception and trust*, it seems reasonable that machines ought to be programmed to trust* different people to different degrees. The interactions between people and machines will also become more complicated if machines become increasingly deceptive themselves and if the machines become increasingly sensitive to human deceptions. If we cannot develop AAs that can appropriately accommodate benevolent and beneficial deceptions in their trust* relationships, then perhaps Joanna Bryson's has a stronger argument that we ought not build such sophistications into AAs (2012).

5 Conclusions

Deception and trust, loaded philosophic terms, are live issues in cyberspace. Increasingly, some deceptions are being realized by machine masquerades. Whether to implement a deception, which implies success at LoAD and misapprehension by users at LoAU, is up to the developers of software at LoAD. These decisions should never be taken lightly. The possible trust/deception scenarios increase in complexity and ethical significance as machines become increasingly sophisticated and capable. We advise caution whenever considering programming deception into an AA.

References

Bryson, J. (2012) Patiency is Not a Virtue: Suggestions for Co-Constructing an Ethical Framework Including Intelligent Artefacts, presented at the Symposium on the Machine Question: AI, Ethics and Moral Responsibility, part of the AISB/IACAP World Congress 2012 (2-6 July 2012), Birmingham, UK.

Carson, Thomas L. (2009) Lying, Deception and Related Concept, The Philosophy of Deception. ed Clancy Martin, New York: Oxford University Press, pp 153-187.

Floridi, L. (2008) The method of levels of abstraction. *Minds and Machines*, 18:303-329. doi:10.0007/s11023-008-9113-7.

Floridi, L. (2013) The Ethics of Information. Oxford University Press.

Grodzinsky, F., Miller K., and Wolf, M. J. (2009) Why Turing Shouldn't Have to Guess. *Asia-Pacific Computing and Philosophy Conference* (Oct. 1-2, 2009) Tokyo.

Grodzinsky, F., Miller K., and Wolf, M. J. (2011) "Developing artificial agents worthy of trust: 'Would you buy a used car from this artificial agent?'" Journal Of Ethics and Information Technology, vol. 13 no 1 Springer Netherlands, March, 2011, pp 17-27.

Grodzinsky, F., Miller K., and Wolf, M. J. (2013) "Automated Deceptions, Benevolent and Otherwise" Thirteenth Ethicomp (June, 2013).

Lynch, Michael P. (2009) Deception and the Nature of Truth, *The Philosophy of Deception.* ed. Clancy Martin, New York: Oxford University Press, pp 188-200.

Mcknight, D. H. and Chervany, N. L. (1996) The Meanings of Trust. http://www.misrc.umn.edu/workingpapers/fullPapers/1996/9604_040100.pdf (accessed March 3, 2013).

Solomon, Robert C. (2009) Self, Deception, and Self-Deception in Philosophy, *The Philosophy of Deception* ed. Clancy Martin. New York: Oxford University Press, pp 15-36.

Strudler, Alan, (2009) Deception and Trust, *The Philosophy of Deception.* ed. Clancy Martin, New York: Oxford University Press, pp 139-152.

Taddeo, M. (2009) Defining trust and e-trust: from old theories to new problems. *International Journal of Technology and Human Interaction* **5**, 2, April-June 2009.

Turkle, Sherry (2011) *Alone Together*, New York: Basic Books.

Wilson, Pete. (2011) Computer spots micro clue to lies. http://www.ox.ac.uk/media/science_blog/111123.html (accessed 18 September 2012).