Ethics Inf Technol (2011) 13:17–27
DOI 10.1007/s10676-010-9255-1

ORIGINAL PAPER

Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?"

F. S. Grodzinsky · K. W. Miller · M. J. Wolf

Published online: 5 December 2010

Abstract There is a growing literature on the concept of *e-trust* and on the feasibility and advisability of "trusting" artificial agents. In this paper we present an object-oriented model for thinking about trust in both face-to-face and digitally mediated environments. We review important recent contributions to this literature regarding e-trust in conjunction with presenting our model. We identify three important types of trust interactions and examine trust from the perspective of a software *developer*. Too often, the primary focus of research in this area has been on the artificial agents and the humans they may encounter after they are deployed. We contend that the humans who design, implement, and deploy the artificial agents are crucial to any discussion of e-trust and to understanding the distinctions among the concepts of trust, e-trust and face-to-face trust.

Keywords Artificial agents · Trust · E-trust · Electronic trust · Modeling trust

A preliminary version of this paper was presented at CEPE 2009. See Grodzinsky et al. (2009).

F. S. Grodzinsky
Sacred Heart University, 5151 Park Avenue, Fairfield, CT, USA
e-mail: grodzinskyf@sacredheart.edu

K. W. Miller
University of Illinois Springfield, One University Plaza,
Springfield, IL, USA
e-mail: miller.keith@uis.edu

M. J. Wolf (✉)
Bemidji State University, 1500 Birchmont Drive NE, #23,
Bemidji, MN, USA
e-mail: mjwolf@bemidjistate.edu

Introduction

In her book *Lying*, Bok writes, "*Whatever* matters to human beings, trust is the atmosphere in which it thrives" (1978). Baier opens her article "Trust and Antitrust" with this same quote (1986). Both the book and the article are about trust among humans. A central purpose of this paper is to develop a model for describing and discussing trust between humans and artificial agents (AAs) and trust among AAs. That is, we provide a mechanism for compartmentalizing features of trust that are present when trust is between two humans, between a human and an artificial agent, and between two artificial agents.

This leads to a series of questions for the software developer who is deciding how to model trust in an artificial agent. What parameters should be in place when the interaction is human to artificial agent (H → AA), artificial agent to human (AA → H), or artificial agent to artificial agent (AA ← → AA)? Is the word "trust" appropriate when discussing interactions that include, sometimes exclusively, AAs? If so, what should AA developers do to create trust in these environments? All three types of interactions, H → AA, AA → H, and AA ← → AA, present different challenges to the developer of an AA, not the least of which is exactly what he/she is trying to model.

We propose the following model, based on object-oriented software design, for thinking about trust to help answer the questions posed above. The model, in Fig. 1, is constructed on the premise that there is an overarching notion of TRUST with attributes of trust found in both face-to-face (f2f) environments and in electronic (e) environments. In this model, f2f-trust is the type of trust that we most commonly associate with human-to-human interactions that occur in a physical space where the participants might possibly touch ("trust with touch"). E-trust is trust

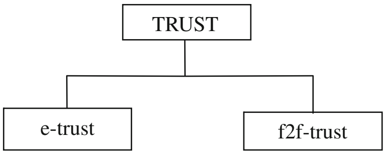

Fig. 1 TRUST includes face-to-face trust and electronically-mediated trust

that develops in a digitally mediated environment ("trust with*out* touch") (Zheng et al. 2002). The diagram suggests that as we identify attributes of each of the two sub-types of trust (f2f-trust and e-trust) there will be some attributes that are common to both and some that are unique to each case. On the one hand, in both digital and physical environments, there are often social norms and references from others that influence a trust relationship. We ascribe these shared attributes to the TRUST superclass. On the other hand, f2f-trust requires physical closeness and digitally mediated trust relationships do not. Thus, physical proximity is an attribute of f2f-trust in our model. We contend that e-trust and f2f-trust are different in ways that are ethically significant.

Because "trust" has been traditionally defined in terms of humans trusting other humans, most often based on face-to-face encounters, we will use the notation "TRUST" in this paper to refer to a broader, more abstract view of the notion of trust. Thus TRUST will include traditional, face-to-face trust between humans, and TRUST will also include electronically mediated relationships and relationships that include artificial agents.

Humans trusting humans (H ← → H) in a face-to-face environment is the source of our primary experience about trust in general. It is certainly an instance of f2f-trust and, thus, TRUST.

In the remainder of the paper, we review some recent work on trust, e-trust, and social trust. We then conceptualize e-trust for developers of AAs by examining instantiations of trust interactions, identifying important attributes of trust for placement in our model, presenting explicit TRUST attributes for those who develop AA's, and finally expanding our initial model with added granularity.

Recent analyses of trust

Trust has been of scholarly interest for a long time. McKnight and Chervany have done extensive research on the different meanings of trust that scholars have used in their research and tied those meanings to common usage of the word "trust" (1996). Their work and the work they analyzed is most directly concerned with H ← → H f2f-trust as we have identified it above. They identified at least six different major definitions of trust that scholars have used in their research. Some recent empirical research has focused on e-trust from an e-commerce perspective. Much of the literature regarding trust in on-line environments is focused on the different influences on trust e-consumers develop in different trust contexts.[1] In this paper we develop a model for trust that encompasses all of these notions of trust and e-trust, allows for the inclusion of other types of e-trust, and gives a structure that is helpful in clarifying analysis.

One particular type of e-trust that we focus on in the model is e-trust that develops with respect to artificial agents. Taddeo has laid the groundwork for the development of our model that supports analysis of various aspects of trust in all of its forms (2009, 2010).

Principles of TRUST

Taddeo analyzes several different definitions of trust and e-trust that have been developed in the past 30 years and presents several problems that remain (2009).[2] Despite the remaining controversies and questions that Taddeo identifies, we require at least an outline of trust and e-trust to accomplish our goal of ethical advice to software developers involved in AA projects. To that end and following Taddeo's analysis, we adopt the following principles about TRUST, the broad idea defined above:

1. **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.
2. **TRUST is a decision by *a* to delegate to *b* some aspect of importance to *a* in achieving a goal.** Note: We rely on the notion that an artificial entity ***a*** includes "decisions" (implemented by, for example, IF/THEN/ELSE statements), and we assume that ***a***'s decisions are designed and implemented with the assumption that there is a high probability that ***b*** will behave as expected.
3. **TRUST involves risk; the less information the trustor *a* has about the trustee *b*, the higher the risk and the more trust is required.** Note: This is true for both artificial and human entities. In AAs, we expect that risk and trust are quantified or at least categorized explicitly; in humans, we do not expect that this proportionality is measured with mathematical precision. In either case, the lack of information might

[1] See the overview article by Benbasat et al. (2008) for a list of influences and reflections on state of the art research in this area.

[2] We will not replay the arguments behind this analysis here; interested readers should see Taddeo's paper, as well as ideas she criticizes in Luhmann (1979), Gambetta (1998), Nissenbaum (2001), Tuomela and Hofmann (2003), Weckert (2005).

depend on the fact that the trustee's behavior is never fully predictable.

4. **The trustor *a* has the expectation of gain by trusting the trustee *b*.** Note: With respect to AAs, "expectation of gain" may refer to the expectation of the AA's designer in moving toward a particular goal, or it may refer to an explicit expression in the source code that identifies this expected gain, or both.
5. **The trustee *b* may or may not be aware that trustor *a* trusts *b*.** Note: If ***b*** is human, circumstances may have prevented ***b*** from knowing that ***a*** trusts ***b***. The same is true if **b** is an AA, but there is also some possibility that an AA trustee ***b*** may not even be capable of "knowing" anything in the traditional human sense.
6. **Positive outcomes when *a* trusts *b* encourage *a* to continue trusting b.** Note: If *a* is an AA, this cycle of TRUST → good outcome → more TRUST could be explicit in the design and implementation of the AA, or else it could be implicit in data relationships, as in a neural net.

We contend that these principles belong to any subset of TRUST be it e-trust or f2f-trust, and therefore, properly belong to the TRUST superclass at the top of our hierarchy.

E-trust

While there are a variety of definitions of e-trust, most require the trust relationship to be digitally mediated, i.e., the *mode* of communication for the trust relationship is digitally mediated. An example is a relationship that occurs exclusively over the Internet. We might assume that all interactions with AAs are e-trust interactions, however, a different example, a physical encounter with a robot, demonstrates that this is not always the case. On the one hand, if a robot encounters a human within physical proximity and uses analog means (like sound or light), not digital electronic signals, to communicate, it will be, according to our definition, "face to face" communication even if the robot does not have a traditional humanoid "face." Any TRUST interactions developed in this scenario are instances of f2f-trust. On the other hand, if the robot is using a digital mode of communication, say telephone or teleconference, then the encounter and subsequent TRUST interactions are instances of e-trust. The digital media mediating e-trust relationships bring their own risks, different risks than those present in f2f-trust relationships.[3]

Taddeo presents an analysis of how to build an assessment model to evaluate e-trust in distributive systems (2010) that involves exclusively AA ← → AA type interactions. She bases this model on the assumptions that fully rational AAs can be designed[4] and that such an agent can choose the best option on whether to trust based on specific information and the agent's goal. We have an immediate concern at this point because there is no guarantee that the developer of an AA has been successful at creating an AA whose behavior can be fairly characterized as "rational." Furthermore, if the AA is capable of changing its internal program[5] after deployment, then its "rationality" is more suspect.

Despite our reservations, we still think a logical next question is important: how can an artificial agent be programmed to behave in a manner similar to how humans behave when humans report that they have learned to e-trust someone or something? In Taddeo's analysis, an AA can measure another entity's trustworthiness according to the ratio of successful actions divided by the total number of actions necessary to achieve a similar goal. Note that the e-trust is not a direct property of the entity, but rather a perceived probability that the entity will accomplish a particular goal. Any entity whose past success rate on a particular act is above a designated threshold would be deemed trustworthy (Taddeo 2010). Entities whose measured performance is below the threshold would be deemed risky or "untrustworthy."

Clearly, a significant risk analysis is required when a threshold is determined. In one scenario, the developers of an AA could set the threshold before the AA is deployed, and it would remain at this value after deployment. In a significantly different scenario, the developers could establish an initial value for the threshold, and then the AA could adjust the threshold after deployment based on events that occur during the AA's interactions with other entities. No matter how or when the threshold is set, the ultimate goal is to have the benefits outweigh the costs of committing to an e-trust relationship.

Taddeo points out that if the trustor has sufficient confidence in the trustee, then the trustor will let the trustee act without supervision, increasing the benefit of the e-trust relationship for the trustor. If the AA trustee has a fixed program, it is reasonable for the trustor to have higher confidence than if the AA trustee can change its program after deployment. Since there is currently little information publicly available about AAs that might be encountered, potential e-trust partners do not normally know anything about the programming of the AA or even about the programming of AAs in general. It may also be difficult to tell whether a potential e-trust partner is human or artificial. In such a situation, AAs necessarily rely on reputation and

[3] In the information systems literature this type of risk is referred to as channel, web, or Internet risk.

[4] See Floridi and Sanders (2004).

[5] See Grodzinsky et al. (2008).

past performance in making decisions about e-trust relationships.[6]

Social Trust

Coming at e-trust from essentially the other extreme, Durante explores trust in terms of a socio-cognitive model of limited rationality for human to human (H ← → H) interaction (2008). His work follows earlier work on the socio-cognitive model done by Castlefranchi and Falcone.[7] This model "is aimed at coping with the uncertainty of what remains beyond control. The idea of trust offers us some insightful elements to reduce uncertainty within cooperative relations" (Castlefranchi and Falcone 2008). Durante also explores the role of delegation in trust relationships and how we move from "control trust" in technology where mechanisms define the trustworthiness of a system to "perceived trust" where trust is based on the confidence that the trustor has in the trustee (Castlefranchi and Falcone 2008). While Durante is primarily investigating online cooperation in terms commons-based peer production among humans, today and in the future there will be a growing possibility that a user might be dealing with AAs and not other humans. Should our notion of e-trust change when we don't know whether the entity we are dealing with is human? Do we consciously delegate responsibility to an AA because we e-trust it, or is this an unconscious decision that we make when we delegate responsibility to our computer? In addition, Durante's focus is on cooperation, but there are times when the trustee is simply facilitating the realization of a goal of the trustor.

For example, when we use Google to search the web, we rely on Google to return a list of sites relevant to our search terms, but most of the time, we don't make a conscious decision to place our trust in Google. At one level of abstraction, the user and Google share the goal of an information exchange about sites relevant to the search terms; at another level of abstraction, Google has no information about the ultimate goal of the entity doing the search. Note that AAs do Google searches, too. No matter who or what is doing a Google search, the searching entity may have goals that Google neither knows about, nor necessarily would approve of. For example, Howe and Nissenbaum's TrackMeNot software does random Google searches to obscure from Google information about the human user's actual searches (2009).

E-Trust for AA developers

The history of humans trusting (and not trusting) humans offers a starting point to our discussion. We humans trust (and don't trust) each other to different degrees. We have some trust issues that are conscious and well thought out; for example, trust is an explicit issue when we vote in elections. We have some trust issues that appear instinctual or at least "decided" before we can articulate logical reasons; for example, a child usually trusts its parents. We have some trust decisions that are initially thought out, but then become habitual; for example, "Should I buy from this website?" These different approaches to trust–largely automatic, explicitly deliberate, and shifting from deliberate to automatic over time–are all relevant to how a software developer must approach designing an AA to interact with humans and with other AAs.

Our first attempt at conceptualizing TRUST that involves AAs is to apply to our model to whatever trust framework a person has for dealing with people. This strategy has the strength of being "species-independent," action centered rather than agent centered, much like Taddeo's approach. That is, we trust individual entities to the degree that their actions warrant that trust, regardless of whether the entities are carbon-based humans or silicon-based AAs. Modeling human trust in AAs after H ← → H f2f-trust would be one way to move beyond what Gunkel describes as the "anthropocentrism" of traditional moral theory (2007). However, the elegant simplicity of this approach may not be appropriate; humans and AAs are not identical, and therefore our approach to applying the TRUST model should take into account those differences where possible. (Identifying when we are dealing with a human and when we are dealing with an AA can itself be problematic, as we will discuss later.) No matter what approach is taken to trusting AAs (or not), humans need to be aware that some AAs will fail in terms of violating trust, in ways similar to those that some humans we choose to trust ultimately violate our trust. Taddeo states that a crucial issue for information and communication technologies is the management of trust, developing parameters for trust to emerge and then finding the methods of assessment (2010).

We use the definition of artificial agents given by Grodzinsky et al.: an ''artificial agent'' is a nonhuman entity that is autonomous, interacts with its environment

[6] The idea that AAs "make decisions" is philosophically controversial; we are not claiming here that AAs make decisions in a way that is identical or comparable to the way humans make decisions. Instead, we are only positing that the program running inside an AA will take different courses of actions based on decision control structures and the state of the computation at a given moment. Although such issues as whether AAs have free will are fascinating and are related to trust, we explicitly are not exploring those issues in this paper.

[7] See Castelfranchi and Falcone and the work of the Unit of AI Cognitive Modelling and Interaction, National Research Council Institute of Psychology, Rome, Italy (2001).

and adapts itself as a function of its internal state and its interaction with the environment (2008). There are numerous objections to this (or any other) definition of AAs. One important objection is that an entity might act as an agent and not have the ability to adapt itself. Although it is a possibility, the adaptable AA is the far more interesting case in ethical analysis. We note that the adaptation could be as simple as changing the value of a variable or as complex as "self-modifying code," by which an AA can change its programming after deployment.

Finally, as background to several of the issues we will consider, we assert that at some point in the future it will not always be easy or convenient (due to lack of direct information about the entity) during an interaction to discern whether an entity is an AA. This is already true for some small subset of interactions mediated by computers. (For example, AAs can write emails that appear to be written by a particular human to a different human.) We expect that soon some phone interactions will be done by AAs in a way that will be hard to distinguish from phone interactions between humans. The use of humanoid robots that are indistinguishable from humans in personal interactions is further away in time, but we do not foresee any decisive reasons why such future developments are impossible. For this reason, unless otherwise noted, our discussion about AAs interacting with humans and with each other include the possibility that the participants are physically proximate, are communicating via a phone, are interacting over the Internet in real time, or are interacting asynchronously over the Internet.

H → AA

Our interest in a human trusting an AA is not merely in describing it. It is vitally important that we explore the issue of whether trusting AAs is a strategy that is likely to lead to positive results. The most straightforward analysis of the issue leads to the somewhat unsatisfying answer "it depends on the AA in question."

There is a revealing example of TRUST that started in the early days of the web: the voluntary standard called the "Robot Exclusion Protocol." By placing a text file called "robots.txt" in a top directory containing web files, the owner requests that all files contained in subdirectories be excluded from all searches done by an automated web searching bot. According to a 2007 study, more than 38% of the websites they examined included a robots.txt file (Sun et al. 2007). Subsequent research reported 2.2 million robots.txt files (Sun et al. 2008). These numbers document an enormous demonstration of human trust in AAs: that at least some web crawling bots (which fit our definition of an AA) will honor this voluntary protocol. Interestingly, some web page owners have sought to shame people who deploy web bots that violate the Robot Exclusion Protocol.[8] The existence of millions of robots.txt files and the attempt at retribution against AAs and their developers who do not honor the voluntary protocol are a singular and contemporary example of TRUST relationships between humans on the one hand and AAs and their developers on the other hand.

What was there about the web-crawling bots in this example that inspired (or at least allowed) TRUST? And, in general, what characteristics of an AA will lead to it being worthy of human trust? Even if the Robot Exclusion Protocol is now ignored by some AA developers, we assume that at some time in the past at least, humans building websites had an expectation that web bots would honor the protocol. (That expectation may or may not have been reasonable, but the fact that so many people bother to include a robots.txt file in their web sites is evidence that the expectation existed.) This may be taken as an example of humans transferring their Trusting Belief (McKnight and Chervany 1996, p. 33) from a H ← → H f2f-trust situation to a H ← → H e-trust situation (website owners trusting web bot programmers), and then to a H → AA e-trust situation. McKnight and Chervany identified benevolence, honesty, competence and predictability as attributes of trust studied by empirical researchers (1996). As we look at the question of trusting AAs from the perspective of the AA developer, predictability is an important part of trust. When dealing with an AA, competence manifests itself as reliability. In our analysis, benevolence and honesty are not important traits of the AA (although they may be important traits of the AA designer). Instead, we introduce the concepts of identity and transparency as important attributes affecting human trust of AAs.

Predictability

Predictability is an attribute from which to draw important distinctions between humans and AAs. AAs are distinct from humans in the sense that we expect that they are capable of changing much faster than humans. Also, the discrete nature of their software increases the likelihood of abrupt and dramatic changes in AAs; in humans, we expect slower, more gradual changes in processes that at least appear to follow laws described with continuous values and mathematics. Some scholars contend that human behavior tends to be modeled more accurately with continuous mathematics than with discrete mathematics, although not all scholars agree (Miller 1988). Because software moves at speeds that are beyond the perception of humans, AAs can go through a dramatic self-modification process multiple times during a relatively slow interaction with a

[8] See Kloth (2009).

human. This sort of change can be disruptive to any existing TRUST relationship that relies on predictability stemming from past experience with that AA. Programmers of AAs could develop an AA to simulate the slower, more gradual changes of humans; however, in any given AA, that programming intent might be thwarted by software faults or unanticipated situations.

Distinguishing between types of implementations of AAs is important here. The simplest situation is when the AA is software that is run on a computer controlled by the human in the e-trust relationship. When the hardware and the software of this computer are completely secure and the software does not modify its own code, the human can be more confident that any e-trust that has been established is still valid.

A slightly more complicated case occurs when numerous people have developed a trust relationship with a single instance of an AA. Without a mechanism by which the humans are alerted that the AA's programming has changed, there is the potential that the TRUST knowledge held by the human can be exploited by the AA. These changes can come about either through self-modification or through developer-directed upgrades. We are considering here AAs that can change future behaviors based on the effects of past actions. When adaptation can obscure the initial design, the behavior of the AA is far less predictable than when the possible adaptations are strictly limited. Developers need to consider the impact of these changes on everyone who has developed a TRUST relationship with the AA. The fundamental normative question is whether the people involved should be informed in some way that the AA has changed. A technical question is whether there will always be an effective method for doing this notification.

Identity

These observations about predictability raise another important issue: TRUST may be tied to whether the human can identify the actual AA he/she is working with. Building TRUST over multiple transactions with an AA requires that the human be in a position to identify the AA. This raises an important ethical consideration for AA developers. At what point does an AA stop being the original AA and become a different one? At what point has the AA changed so significantly, that any reasonable human who has interacted with the AA in the past should no longer rely on the TRUST history that has been built up? Answers to these questions are complicated by the possibility that some humans may have had multiple recent interactions with recent versions of the AA and other interactions may have been with much older versions. If a human is unable to identify the AA or its version history, then that human is at increased risk if he/she trusts the AA.

There is strong motivation for developers to encourage humans to trust the AAs they develop. When humans do not trust an AA, it is likely that it will be unused and perhaps avoided by humans. In a commercial setting, such avoidance would be disastrous. There are different strategies that a developer could use to encourage TRUST, including deception. But when a developer chooses to honestly earn the TRUST of humans interacting with an AA, programming transparency about versions and identity are one strategy for making that AA appear worthy of TRUST.

Reliability

Even in the simplest AA interactions, there is always an element of risk. The saying "no one is perfect" originally referred to humans, and it is no less applicable to AAs. Artificial agents are becoming increasingly sophisticated, and the software size and complexity necessary for that sophistication increases the likelihood of software faults (Clarke and Wing 1996). The problems of validation and verification for such systems are particularly acute for adaptive systems such as neural nets (Schumann and Nelson 2002). The potential for making an individual AA more *useful* to humans encourages increasing its complexity. The lure of making humans more *comfortable* with AAs also encourages this complexity, at least when "more human-like" is equated with increased comfort (DiSalvo et al. 2002). However, it is important to note the likely increase of risk (and the decrease of trustworthiness) that increased complexity can entail.

Some AAs are explicitly designed to harm humans, such as lethal robots used in warfare. We are not focusing on such AAs here. Instead we are exploring potential harms that are unintentional.

The relationship between AA software reliability and human TRUST in AAs is part of a much larger issue of the impact of software quality on humans (Wolf and Grodzinsky 2006). We cannot explore this issue in any depth here, but note that the many arguments for simplifying software artifacts in order to increase their reliability are particularly relevant to the issue of H → AA TRUST. As AAs proliferate, and as they become increasingly responsible for important aspects of human lives, the reliability of those AAs is a direct responsibility of the AA developers. The seemingly autonomous behavior of sophisticated AAs may mask the role of the developers who launched the AA; but this does *not* absolve the AA developers from their responsibilities for the immediate and the future consequences of the AA's deployment.

Transparency

Whenever data mining is used, ethical issues arise. These issues can become particularly complex if data mining is used to select and weigh indicators (Fule and Roddick 2004). But we assert that these issues are particularly troubling when an AA collects and analyzes data mining results and then applies the result of that analysis to influence its own actions. Such actions can have significant effects on humans without possible interventions from other humans. It is one thing to program an AA to discern patterns based on fixed criteria in data available to the AA after deployment; it is quite another to allow the AA to adjust or augment those criteria based on its own analysis of data-mined information after deployment. Imagine being denied a loan because the decision was made by an AA using data mining information collected during a deep recession. In the best case, the human might be able to appeal the decision; in the worst case, the AA might have the final say, based on a neural net "analysis" that would be difficult if not impossible to explain to a human.[9]

Transparency at best and traceability at least, is a theme we raise in this section, and reprise in the next. If humans are to risk TRUST in AAs, then AA developers should produce systems whose criteria and processes for making decisions are accessible to humans. If these systems' decision making processes are obscure or hidden, humans are less likely to trust AAs over the long run, and we assert that humans *should* not trust such systems. The formal nature of software makes transparency possible for AAs in a way that is not possible for humans. In the area of transparency, it may be possible (while acknowledging the potential for deception) to exhibit trustworthiness in AAs more readily than in humans.

AA → H

In this section we consider an AA "trusting a human." We do not want to defend the claim that a computer program "experiences TRUST" in a way that is identical to humans. Instead, we want to explore (without a protracted debate about the nature of an AA's "experience") the *behavior* of an AA that would appear to an outside observer to be based on a relationship of TRUST. In this respect we are using levels of abstraction in a way consistent with Floridi and Sanders (2004), but without conceding that an AA should be declared a moral agent.

Measurement

In the previous section we explored humans making judgments about AAs and their trustworthiness. In this section, we explore the process and implications of AAs that measure human trustworthiness and act accordingly. For example, an effective software agent used to buy goods offered on the web should not merely look for the lowest price. Although price is surely a factor, the reliability of the seller is also important. According to our definitions, such an AA is deciding which entity can be *e-trusted* sufficiently to risk a purchase.

When an AA developer is programming an AA to perform a task that requires such a "judgment," what indicators can the AA use to decide among possible trading partners? And among the possible indicators, which indicators *should* be selected by the AA developers? Issues of justice and fairness are clearly at stake here. In this section (as well as the next), we note that a software developer dealing with these AA → H trust behaviors will have to make explicit notions about TRUST that may be vague and amorphous in human trust relationships. For example, various scholars have identified a trait referred to as Dispositional Trust (McKnight and Chervany 1996, p. 37) that develops over people's lifetimes and is an important element in establishing trust; but software developers will be hard pressed to program such notions, at least with currently available AI techniques.

As AA developers create, deploy and gain experience with AAs that explicitly model AA → H trust, they will be able to collect and analyze data about which AA decisions and protocols for trust succeed and which fail according to some objective criteria about what constitutes a successful AA → H trust relationship. By analyzing this data, psychologists and philosophers may be able to make new hypotheses about the mechanisms and patterns of AA trust, which may offer insights into human trust.

AA ← → AA

When the need for TRUST arises in an AA to AA interaction, what criteria should AA_1 use to decide if AA_2 is trustworthy? As in Taddeo's model where the criteria are based on past similar experience (2010), the details of the criteria will necessarily be application-specific, but we present some general principles to guide AA developers producing AAs that will need to be trustors or trustees in their interactions with other AAs.

As in the H → AA section, we contend that transparency is vital when determining the trustworthiness of decision-making procedures and sources of data. Transparency can be difficult when a commercial AA's decision-making details constitute a trade secret; even in such cases,

[9] There are both theoretical and practical reasons why neural net decisions are unlikely to be easily explained to humans. For example, systems that can give a comprehensible explanation to a human of why a decision was reached are far more resource intensive than systems that are less expressive about their reasoning (Greiner et al. 2001).

the AA's details could be made known to designated third parties (regulators or consultants) dedicated to keeping AA interactions fair. In addition to this overall transparency, individual decisions should be traceable, so that disputed decisions in AA ← → AA interactions can be investigated ex-post facto for purposes of undoing any injustices and for analyzing what went wrong in order to improve future interactions.

Explicit TRUST criteria

At least initially, AA developers must necessarily make explicit the criteria the AA will use in making decisions. (If the AA can self-modify, those criteria can change.) This formalizing activity is a difficult one requiring delicate judgments and ethical sensitivity. It is also an opportunity to explore methods of making such decisions justly and efficiently. When AA developers take these ethical challenges seriously, their deliberations, the resulting AA programs, and data about the results of their programming as the AAs are deployed and used are likely to be useful to philosophical and political debates about both human and AA decision-making policies.[10]

Testing

Assessment methods and risk analyses need to be developed to evaluate parameters and the trustworthiness of AAs. While developers of unmodifiable AAs can set their parameters and test them, how can anything but initial parameters be relied upon in a modifiable AA? The testing of modifiable AAs is far more complicated than the testing of unmodifiable AAs. Unless the AA includes safeguards that are effectively shielded from future modifications, the possibility of modifications in response to unforeseeable future circumstances make testing of modifiable AAs at the very least impractical and probably impossible in any reasonable amount of time. Even "protected" safeguards may be vulnerable to future modifications (Grodzinsky et al. 2008).

Testing any software of even modest complexity is a major challenge in software engineering; software that can modify itself makes that challenge unmanageable. We contend that self-modifications after deployment that could affect an AA's behavior should be severely limited by effective and well designed safeguards; it may be advisable to avoid self-modifying AAs altogether because of the inherent risks. Great caution is required in this area if humans (and AAs for that matter) are expected to rely on the trustworthiness of AAs.

Expanding the model[11]

The analysis of the variety of types of interactions and the distinctions among them suggest a richer explanatory model is needed. We have identified three main aspects of trust interactions: the mode of communication, the type of trustor and the type of trustee. That is, is the communication mode electronic or face-to-face, is the trustor (X) human or artificial and is the trustee (Y) human or artificial?

As before, we identify electronic communication with an "E." To simplify notation and in recognition of the possibility that the entities involved may or may not have faces and may not be persons, we identify face-to-face communication with a "P," where "P" stands for physical, or proximate. This model is more granular with eight distinct subclasses of the TRUST class.

The first stage of analyzing a particular, contextualized instance of TRUST is to identify which of the eight possible subclasses it belongs to. This stage is similar to defining a subclass of a superclass. In stage 2, the practitioner should answer two more questions: What is the socio-technical context of the *relationship* between X and Y? What is the socio-technical context of the *communication* between X and Y? These questions are decidedly not binary. Indeed, the answers can be arbitrarily complex. This stage is similar to instantiating an object from a subclass and adding information to that instantiated object.

To facilitate exposition, we use "XYZ-TRUST" to represent our superclass TRUST. The intent is that X and Y classify the trustor and trustee, respectively, as either human (H) or artificial (A); and that Z specifies whether the communication is physical or electronic. Since X, Y, and Z are binary, there are eight possible combinations for XYZ. Each of these eight combinations is a subclass of our superclass. Figure 2 provides an organization of the subclasses that shows useful relationships among the subclasses.

Note that column 1 is human trusting human, column 2 is human trusting machine, column 3 is machine trusting human, and column 4 is machine trusting machine. The first row requires physical proximity and the second row is electronically mediated. The first column is all human, and the remaining three columns require at least one artificial communication partner. Thus the most traditional idea of 'trust' is the top left corner, and only 1/8 of our expanded model of trust.

[10] The learning that can take place studying AA trust decisions will be facilitated if that software is available widely. Thus, source code available software (including Free Software) will be of particular interest.

[11] This expanded model was first developed in Grodzinsky et al. (2010).

HHP	HAP	AHP	AAP
HHE	HAE	AHE	AAE

Fig. 2 The participants and mode of interaction define 8 distinct subclasses of TRUST

- HHP-trust: traditional notion of human, "face-to-face" trust
- HHE-trust: humans trust each other, but mediated by electronic means
- HAP-trust: human trusts a physically present AA, for example, a robot (no electronic mediation)
- HAE-trust: human trusts an artificial entity (like a web bot) over the Internet
- AHP-trust: an AA, perhaps a robot, trusts a physically present human
- AHE-trust: an AA, perhaps a web bot, trusts a human based on Internet interactions
- AAP-trust: an AA trusts another AA in a physical encounter; because this is P-trust, the AAs might, for example, use sign language
- AAE-trust: an AA trusts another AA electronically, e.g., two web bots communicate via the Internet

The subclasses in Fig. 2 are derived subclasses of the superclass XYZ-TRUST; each subclass has characteristics that distinguish it from any other subclass. In addition, within each derived subclass, each instantiation of that subclass can include a distinctive combination of socio-technical contexts. As a consequence different instantiations of HHP-trust that are within the same subclass may have quite different contexts. For example, imagine these two scenarios, both assumed to happen in physical space, not Cyberspace: a 3-year-old daughter trusts her mother to feed her, and a 40 year old employee trusts her employer to pay her. Both of these scenarios are instantiations of the subclass HHP-trust, and both share characteristics of that subclass; but the two instantiations are otherwise quite different because of the socio-technical context of the particular H–H pair. Notice that within a type of communication, there can be quite different socio-technical contexts; for example, both video Skype communication and email communication are electronic, but they represent communication contexts that differ substantially. Thus the inclusion of a socio-technical context is important both for the entities involved (X and Y) and for the communications (Z) used to establish the trust relationship.

Conclusions

In the previous sections, we used H ← → H, H → AA, AA → H, and AA ← → AA trust interactions and our model of TRUST with its subclasses, e-trust and f2f-trust, to discuss areas of concern that influence modeling trust in AAs. Thus, this model helps conceptualize some of the issues that AA developers must wrestle with as they build the AAs. We offer the expanded model to practitioners as a practical method of approaching the problem. Perhaps the most important observation based on this model is how complex the issue of trust becomes when electronic communication and artificial entities are included in the discussion. When electronics are excluded from the model, then only HHP-trust remains. HHP-trust is itself a huge field, including all the issues of humans trusting (and not trusting) each other on the basis of physical encounters. The expanded model aids analysis when instances of subclasses are paired with socio-technical contexts. Thus the model does not seek to over-simplify the situations we are discussing, but it seeks to organize thinking about these situations in a coherent fashion.

In our analysis, there is a compelling case that the attributes reliability, transparency, predictability, identity, measurement, and testing are properly attributes of the XYZ-TRUST superclass. The implication is that each of these attributes must be addressed in the development of AAs and dealt with in a way that is appropriate for the type of interaction the AA will have.

As AAs become more sophisticated, we expect that it will be increasingly difficult for a human to determine when an entity with which they are interacting is human or artificial. Unless the speed of the interaction is obvious (artificial entities are capable of faster interactions), it may also become commonplace for AAs to interact in a similar way to humans and other AAs; this will be especially true of asynchronous interactions.

When this discrimination problem becomes commonplace (for example, when AAs can routinely pass a Turing Test),[12] there are at least two obvious strategies for both humans and AAs to deal with the situation in which the interaction partner is not known to be human or artificial. One strategy is that humans and AAs adopt a common protocol for interactions, a protocol that does not discriminate between artificial and human partners. This strategy is elegantly simple, but it precludes taking advantage of some possibilities that exist when partners in an interaction know more about each other. For example, an AA ← → AA interaction can occur much faster if the classification of the partner is known. A less technical advantage of knowing is that humans may be more comfortable interacting with a human, or at least knowing as much as possible about the identity of the interaction partner.

The importance of clear thinking about TRUST is directly related to development of more human-like robots

[12] See Miller et al. (2009).

and automated voices because it is assumed that humans will trust robots more when the robots become more like humans (Bruemmer et al. 2004). However, our analysis above suggests that a desire for more human-like characteristics may result in less reliable and less transparent AAs that will ultimately decrease rather than enhance TRUST in AAs.

One useful and distinctive characteristic of humans is our capacity to adapt. However, this adaptability can also lead to capriciousness and unpredictability. Some AA developers are attempting to make AAs more human-like by programming them to be more adaptable to their environment by allowing them to self-modify their programs. We contend that the potential gains of this strategy are not sufficient to justify the enormous risks, especially when the adaptation process is poorly understood by the developer and not easily recognized by humans who have e-trust relationships with the AAs. We prefer that AAs be boringly predictable. We are far more concerned about the trustworthiness of AAs and far less concerned that they mimic human adaptability. In almost all situations, we think that AA developers have an obligation to the safety of the public. That duty should restrict their use of self-modifying code to implement AAs and place limitations on the use of neural nets in AAs.

We are convinced that TRUST will be an increasingly important issue as interactions between humans and AAs increase in frequency and importance. We have attempted to demonstrate that a model-based examination of trust can help organize discussions about different aspects of this complex subject.

References

Baier, A. (Jan, 1986). Trust and antitrust. *Ethics*, *96*(2), 231–260.

Benbasat, I., Gefen, D., & Pavlou, P. A. (Spring, 2008). Special issue: Trust in online environments. *Journal of Management Information Systems, 24*(4), 5–11.

Bok, S. (1978). *Lying* (p. 31n). Pantheon Books: New York.

Bruemmer, D., Few, D., Goodrich, M., Norman, D., Sarkar, N., Scholtz, J., Smart, B., Swinson, M. L., & Yanco, H. (2004). How to trust robots further than we can throw them. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria, April 24–29, 2004). CHI '04 (pp. 1576–1577). ACM, New York.

Castlefranchi, C. & Falcone, R. (2008). Socio-cognitive model of trust: basic ingredients, online at http://istc.cnr.it/T3/trust. Accessed March, 2008.

Clarke, E. M. & Wing, J. M. (Dec, 1996) Formal methods: state of the art and future directions. *ACM Computing Surveys, 28*(4), 626–643.

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). All robots are not created equal: The design and perception of humanoid robot heads. In *Proceedings of the 4th conference on designing interactive systems: processes, practices, methods, and techniques* (London, England, June 25–28, 2002). DIS '02 (pp. 321–326). ACM, New York, NY.

Durante, M. (2008). What model of trust for networked cooperation? Online social trust in the production of common goods (knowledge sharing). In *Ethicomp 2008, conference proceedings* (pp 211–223). University of Pavia, Mantua, Italy.

Grodzinsky, F. S., Miller, K., & Wolf, M. J. (2010) Toward a model of trust and e-trust processes using object-oriented methodologies. In *Ethicomp 2010 Proceedings*, April 14–16, 2010.

Falcone, R. & Castelfranchi, C. (2001) Social trust—A cognitive approach, online at http://istc.cnr.it/T3/trust. Accessed 3/2008.

Floridi, L. & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Fule, P. & Roddick, J. F. (2004) Detecting privacy and ethical sensitivity in data mining results. In *Proceedings of the 27th Australasian conference on computer science—Vol. 26* (Dunedin, New Zealand). Estivill-Castro, Ed. ACM International conference proceeding series (vol. 56, pp. 159–166). Australian Computer Society, Darlinghurst, Australia.

Gambetta, D. (1998) Can we trust trust? In *Trust: Making and breaking cooperative relations*. D. Gambetta, Ed. (pp. 213–238).

Greiner, R., Darken, C., & Santoso, N. I. (Mar, 2001) Efficient reasoning. *ACM Computing Surveys, 33*(1), 1–30.

Grodzinsky, F. S., Miller, K., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*, 115–121.

Grodzinsky, F. S., Miller, K. & Wolf, M. J. (2009) Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?" In *Proceedings of computer ethics, philosophical enquiry conference*, June 26–29, 2009.

Gunkel, D. J. (2007). Thinking otherwise: Ethics, technology and other subjects. *Ethics and Information Technology, 9*, 165–177.

Howe, D. & Nissenbaum, H. (2009) TrackMeNot. http://mrl.nyu.edu/~dhowe/TrackMeNot/, accessed April 7, 2009.

Kloth, R. (2009) List of bad bots. http://www.kloth.net/internet/badbots.php, accessed March 31, 2009.

Luhmann, N. (1979). *Trust and power*. Chichester: Wiley.

Mcknight, D. H., & Chervany, N. L. (1996) *The meanings of trust*. http://misrc.umn.edu/wpaper/WorkingPapers/9604.pdf, accessed June 1, 2010.

Miller, J. (June, 1988) Discrete and continuous models of human information processing: Theoretical distinctions and empirical results, *Acta Psychologica*, *67*(3), 191–257.

Miller, K., Grodzinsky, F., & Wolf, M. J. (2009). Why turing shouldn't have to guess. The Fifth Asia-Pacific Computing and Philosophy Conference, October 1–2, 2009.

Nissenbaum, H. (2001). Securing trust online: Wisdom or oxymoron. *Boston University Law Review, 81*(3), 635–664.

Schumann, J. & Nelson, S. (2002). Toward V&V of neural network based controllers. In D. Garlan, J. Kramer, & A. Wolf (Eds.). *Proceedings of the first workshop on self-healing systems* (Charleston, South Carolina, November 18–19, 2002), WOSS '02 (pp. 67–72). ACM, New York, NY.

Sun, Y., Councill, I. G., & Giles, C. L. (2008). BotSeer: An automated information system for analyzing web robots. In *Proceedings of the 2008 eighth international conference on web engineering—Volume 00* (July 14-18, 2008) (pp. 108–114) International Conference On Web Engineering. IEEE Computer Society, Washington, DC.

Sun, Y., Zhuang, Z., & Giles, C. L., (2007). A large-scale study of robots.txt. In *Proceedings of the 16th international conference on world wide web* (Banff, Alberta, Canada, May 08-12, 2007). WWW'07. (pp. 1123–1124) ACM, New York, NY.

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.

Taddeo, M. (2010) Modeling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines 20*(2), 243–257.

Tuomela, M., & Hofmann, S. (2003). Simulating rational social normative trust, predictive trust, and predictive reliance between agents. *Ethics and Information Technology, 5*(3), 163–176.

Weckert, J. (2005). Trust in cyberspace. In R. J. Cavalier (Ed.), *The impact of the internet on our moral lives* (pp. 95–120). Albany: University of New York Press.

Wolf, M. J. & Grodzinsky, F. S. (2006). Good/fast/cheap: contexts, relationships and professional responsibility during software development. In *Proceedings of the 2006 ACM Symposium on Applied Computing* (Dijon, France, April 23–27, 2006). SAC'06 (pp. 261–266). ACM, New York, NY.

Zheng, J., Veinott, E., Bos, N., Olson, J. S., & Olson, G. M. (2002). Trust without touch: jumpstarting long-distance trust with initial social activities. In *Proceedings of the SIGCHI conference on human factors in computing systems: Changing our world, changing ourselves* (Minneapolis, MN, USA, April 20–25, 2002). CHI'02 (pp. 141–146). ACM, New York, NY.