Philos. Technol. (2015) 28:75–90
DOI 10.1007/s13347-014-0165-8

ORIGINAL PAPER

Levels of Trust in the Context of Machine Ethics

Herman T. Tavani

Received: 16 January 2014 / Accepted: 16 April 2014 / Published online: 3 May 2014

Abstract Are trust relationships involving humans and artificial agents (AAs) possible? This controversial question has become a hotly debated topic in the emerging field of machine ethics. Employing a model of trust advanced by Buechner and Tavani (*Ethics and Information Technology 13*(1):39–51, 2011), I argue that the "short answer" to this question is *yes*. However, I also argue that a more complete and nuanced answer will require us to articulate the various *levels of trust* that are also possible in environments comprising both human agents (HAs) and AAs. In defending this view, I show how James Moor's model for distinguishing four levels of ethical agents in the context of machine ethics (Moor, *IEEE Intelligent Systems 21*(4):18–21, 2006) can help us to develop a framework that differentiates four (loosely corresponding) levels of trust. Via a series of hypothetical scenarios, I illustrate each level of trust involved in HA–AA relationships. Finally, I argue that these levels of trust reflect three key factors or variables: (i) the level of *autonomy* of the individual AAs involved, (ii) the degree of *risk/vulnerability* on the part of the HAs who place their trust in the AAs, and (iii) the kind of *interactions* (direct vs. indirect) that occur between the HAs and AAs in the trust environments.

Keywords Artificial agents · Autonomy · Ethical agents · Machine ethics · Trust

Abbreviations

AA	Artificial agent
AAA	Autonomous artificial agent
AMA	Artificial moral agent
FAAA	Functionally autonomous artificial agent
HA	Human agent

1 Introduction

Research in a relatively new sub-field of information/computer ethics called "machine ethics" (also sometimes referred to as "robo-ethics," "agent ethics," and "bot ethics")

H. T. Tavani (✉)
Department of Philosophy, Rivier University, 420 Main St., Nashua, NH 03062, USA
e-mail: htavani@rivier.edu

has generated a cluster of questions that are metaphysical and epistemological in nature as well as ethical and meta-ethical.[1] In particular, research in this sub-field has raised some core metaphysical questions pertaining to agency (and moral agency) and autonomy, as well as epistemological and ethical/meta-ethical questions affecting trust in connection with artificial agents (AAs).[2] With regard to questions about agency and moral agency vis-à-vis artificial entities, it would seem that many researchers assume that the answers must be framed in ways that are exclusionary or "binary"—e.g., an artificial entity either can *or* cannot qualify as an agent, and (if it can be an agent) either it can or cannot qualify as a *moral* agent (i.e., as an artificial moral agent or AMA). The general consensus among researchers in machine ethics so far seems to be that artificial entities can qualify as AAs of some sort.[3] Regarding the question of whether (some of these) AAs might also qualify (in principle, at least) as AMAs, two opposing camps have formed. However, an alternative scheme for analyzing this question has been proposed by James Moor (2006), who differentiates four levels of "ethical agents" (in his analysis of controversial agency-related issues in machine ethics). In Moor's model, answers to questions about moral agency vis-à-vis AAs need not be framed in ways that require an all-or-nothing (moral) status for the AAs under consideration.[4] I believe that a similar move can also be made in the case of questions involving trust vis-à-vis AAs.

In this essay, I will assume that some artificial entities can indeed qualify as AAs—a position that I have defended elsewhere.[5] I leave open the question of whether an AA could ever be a full AMA. (At a later point in this essay, however, I will argue that some AAs can have "functional autonomy.") But I also argue that even if AAs cannot qualify as full moral agents, they are nonetheless capable of being in trust relationships with human agents (HAs) (a view that is also defended elsewhere[6]). Additionally, and perhaps more importantly for purposes of this essay, I argue that the various levels of

[1] Wallach and Allen (2009) and Anderson and Anderson (2011) use the expression "machine ethics" to describe this relatively new field or sub-field. Others, however, such as Verrugio (2006) and Decker and Gutmann (2012) use the expression "robo-ethics," while Lin et al. (2012) use the phrases "robot ethics" to describe this field.

[2] Whereas I use the expression "AA" in this essay, some researchers in the field of machine ethics use the term "machine" to refer to (computer-generated) artificial entities. Wallach and Allen use the expression "(ro)bot" to capture the broad meaning of AAs or "machines," which, in their view, can be either pure digital entities (such as soft bots) or physical robots.

[3] Some philosophers, however, have questioned whether artificial entities are capable of being genuine "agents" because, in their view, these entities are not capable of *acting* or "performing" an act(ion). While (human) agents are typically viewed as persons that can *act* either on their own behalf or on the behalf of others, some critics suggest that an artificial entity is capable merely of a "doing"; in this scheme, an "act(ing)" is far more complex than a mere "doing." However, I will not further examine this distinction here. (For a fuller discussion of this point, see Himma 2009.)

[4] Just as Moor's conceptual framework involving "logical malleability," "policy vacuums," and "conceptual muddles" (in his seminal essay "What Is Computer Ethics?" (Moor 1985)) provides an alternative and insightful scheme for analyzing a classic debate in computer ethics about whether any ethical issues in computing are "unique" ethical issues, I believe that his framework of four levels of "ethical agents" provides an insightful and alternative scheme for analyzing the current debate in machine ethics about whether AAs can qualify as moral agents.

[5] See, for example, Tavani (2012) and Tavani and Buechner (2014).

[6] See Buechner and Tavani (2011) for a detailed analysis of how "trust relationships" between HAs and AAs are possible within certain contexts—i.e., in the kinds of "zones" that Walker (2006) calls "default trust" and "diffuse-default trust" (described in section 3 of this essay).

trust that are possible in trust relationships between HAs and AAs warrant further elucidation. In developing this point, I also show how we can apply Moor's ethical-agency model in constructing a parallel framework that helps us to better understand the various levels of trust that can be articulated.

2 Moor's Four Levels of "Ethical Agents"

As noted above, arguments have been advanced to defend answers on both sides of the question of whether AAs can satisfy the necessary conditions for being AMAs. For example, Himma (2009) argues that because AAs lack freedom, consciousness, and intentionality, they cannot meet the requirements for moral agency. Johnson (2006) also argues that AAs cannot qualify as moral agents, because they lack freedom; however, she believes that since AAs have "moral efficacy," they can nonetheless qualify as "moral entities." Floridi (2011), on the contrary, claims that *autonomous* AAs, or what he calls "AAAs," can be moral agents because they are (a) "sources of moral action" and (b) can cause moral harm or moral good. However, Moor takes a very different approach in his analysis of this controversial question by focusing on the various kinds of "ethical impacts" that AAs can have.

First, Moor notes that computers (and computer programs) can be viewed as *normative* agents, independently of whether they are also capable of qualifying as ethical agents. Whereas some authors tend to slide back and forth between "ethical/moral agent" and "normative agent" when discussing AAs, Moor, applying a standard distinction in normative ethics, is careful to separate the two categories by differentiating "normative non-moral" and "normative moral" aspects of an AA's impact. In Moor's view, an AA qualifies as a normative agent simply in virtue of whether it can be evaluated in terms of how well or how poorly it performs the task(s) it was designed to do. So, on these grounds, a computer can clearly qualify as an AA in a normative non-moral sense. Consider that many computers are designed for specific purposes and thus can be evaluated normatively—i.e., in terms of how well or how poorly they perform the tasks they are programmed to carry out. For example, Moor considers a case where we might evaluate how well a computer program designed to play chess (such as IBM's Deep Blue program) performs. However, Moor also points out that computers not only have "normative impacts" that are non-moral in nature, but that in some cases, those impacts can be moral or ethical as well. In his scheme, the consequences, and potential consequences, of what Moor calls "ethical agents" are analyzed in terms of four levels: ethical-impact agents, implicit-ethical agents, explicit-ethical agents, and full-ethical agents.

Moor's *ethical-impact agents* represent what he describes as the "weakest" sense of ethical agent in his four-level framework. Nevertheless, these agents carry out actions that have ethical consequences. But these agents are also different from *implicit-ethical agents* in one very important respect: the latter typically have some "ethical considerations" built into their design. Implicit-ethical agents, in Moor's scheme, also "employ some automatic ethical actions for fixed situations." So, there are some significant differences that distinguish the agents occupying the first two levels. At the next level, *explicit-ethical agents* have, or at least act as if they have, "more general principles or rules of ethical conduct that are adjusted and interpreted

to fit various kinds of situations." Finally, *full-ethical agents*—representing the strongest sense of "ethical agent" in Moor's framework—"can make ethical judgments about a wide variety of situations" and in many cases can also "provide some justification for them."

Moor nicely illustrates the first two levels of ethical agents with some specific examples. In the case of an ethical-impact agent, he uses the example of a "robotic camel jockey" (a technology used in Qatar to replace young boys as jockeys, and thus freeing those boys from slavery in the human trafficking business). Moor also provides two examples to illustrate implicit-ethical agents: an airplane's automatic pilot system and an ATM (automatic teller machine). He further notes that both technologies have built-in programming code designed to "prevent harm" from happening; in these two instances, one agent is designed to prevent physical harm to the passengers and crew onboard an airplane, and the other to prevent ATM customers from being short-changed in financial transactions.

Although Moor does not include any specific examples of explicit-ethical agents (as he acknowledges that no AAs may yet qualify for this category), he notes that an AA at this level of "agent" would be able to (a) calculate the best ethical action to take in a specific situation and (b) make decisions when presented with ethical dilemmas. (Perhaps some categories of "carebots" in the near future will succeed in satisfying these criteria.) Finally, Moor describes full-ethical agents as having the kinds of ethical features that we typically attribute to full-blown ethical agents like ourselves (i.e., what Moor describes as "normal human adults"), including the attributes of consciousness and free will. It is also important to note that Moor does not claim that, aside from humans, either explicit- or full-ethical agents exist or that they will be available anytime in the near term. Nevertheless, his distinctions are very helpful, as we try to understand various levels of ethical agency that apply and can potentially apply to AAs.[7] In section 4 of this essay, we show how Moor's framework can be extended to questions pertaining to trust vis-à-vis AAs, and we also show how that model can help us better understand the levels or degrees of trust in the various kinds of trust relationships that are possible between HAs and AAs. First, however, we describe key aspects of the models of trust and autonomy employed in this essay.

3 Trust, AAs, and Functional Autonomy

Some will no doubt question whether it makes sense to talk about trust in the context of AAs; for example, one might argue that trust relationships require (full) moral agency on the part of both the trustor and the trustee, and while HAs are capable of being full moral agents, AAs (at least the AAs of today) are not. As noted earlier, however, I argue that AAs do not necessarily have to qualify as AMAs (or artificial moral agents) to be capable of being in trust relationships with HAs. To see why this is possible, we first need an adequate theory of trust. Unfortunately, many conventional definitions of trust focus mainly or even exclusively on properties such as reliability and

[7] My summary of Moor's model in section 2 of this essay draws substantially from my analysis of his framework in Tavani (2013).

dependability. But definitions of trust based solely on these characteristics will not suffice.[8] Fortunately, some more sophisticated theories of trust, especially with respect to trust in the context of AAs, have been advanced by Taddeo (2010a), Durante (2010), and Grodzinsky et al. (2011).[9] While these theories improve dramatically upon some of the more conventional definitions of trust found in other works, I believe that a different kind of theory is needed to appreciate some of the more nuanced aspects of trust that correspond to the various levels of AAs (in Moor's framework). So, I expand on a model of trust that I have defended elsewhere, which views trust as a (moral or non-moral) normative relationship affecting two agents—A and B—in which "*A* has the disposition to *normatively* expect that *B* will do such and such responsibly."[10] Of course, one might ask how the kinds of normative expectations described in this model—i.e., the kinds of expectations and obligations that we, as HAs, typically have toward one another—can apply in the case of AAs.[11] To understand how this is possible, consider that a trust relationship typically involves agents (both human and non-human[12]) acting within a certain kind of context or environment—one that Walker (2006) calls a "zone of default trust."

Walker notes that in these various zones, which comprise our various "communities," we come to know "what to expect" from others and "whom to trust." In fact, we develop a sense of safety, she argues, because trust is "the default" in these zones. Walker also notes that the trust relationship can even be "diffuse" in these various zones, as it can affect our relationships with large organizations as well as with specific persons. Expanding on Walker's insight, Buechner and the present author (2011) argue that these zones of diffuse-default trust can consist not only of individual HAs and aggregates of HAs, but can also include AAs (such as intelligent software agents and physical robots) as well. We have also argued that trust relationships between the agents (both human and artificial) in either a default or a diffuse-default trust zone need not always involve direct contact between the agents, but alternatively can be realized

[8] For example, consider that while I may depend on my car to start tomorrow, and I while may rely on it for getting to work tomorrow, it would be a bit odd for me to say that I trust my car to start tomorrow.

[9] All three authors provide sophisticated theories of trust in connection with AAs. See, for example, Taddeo's model of a "non-psychological approach" to trust, which rests on a "Kantian regulative ideal of rational agent"; Durante's "socio-cognitive" approach to trust in the context of multi-agent systems; and the "object-oriented" model of trust in Grodzinsky et al. Unfortunately, I am unable to analyze, or even summarize, those trust theories here, because of space limitations. Interested readers will likely want to examine in detail the models of trust presented in each of these three works. The interested reader may also wish to consult Taddeo (2009) for an excellent discussion of the main theories of trust and e-trust (affecting digital environments) that have been published during the past 20 or so years.

[10] See Buechner and Tavani (2011) for a fuller account of this model, which has five distinct conditions or requirements that show how one acquires a "disposition to trust" based on certain normative expectations. (It is worth noting, however, that the sense of "disposition" used in this model refers to the mental state of the trustor and thus should not be confused with behavioristic accounts of dispositions.) Some alternative models that also stress the normative expectations that underlie a trust relation involving HAs are described in Baier (1986) and McLeod (2011).

[11] Note that because this essay is mainly concerned with normative issues, including "obligation," that pertain to "ethical trust," I do not explicitly examine the conditions required for "epistemic trust" involving AAs. For an excellent discussion of aspects of epistemic trust and AAs, see Simon (2010).

[12] The domain of "non-human agents" includes aggregates of individual HAs, such as organizations, institutions, and corporations, as well as (purely) computer-generated AAs. Consider that we frequently enter into trust environments that include not only HAs but also non-human agents (such as financial institutions and corporations that manufacture automobiles).

indirectly—i.e., via the relations and normative expectations defining a diffuse-default environment or *context*.[13]

While HAs can enter into trust relationships with a wide range of AAs, via (default and diffuse-default) zones of trust, I believe that it is important to differentiate both the:

(a) *Kinds of AAs* with which HAs can interact in various trust zones;
(b) *Strength of the trust relationships* that are possible (between HAs and AAs comprising these zones).[14]

Regarding (a), many different kinds of AAs—including distributed AAs and multi-agent systems, some of which may consist merely of software programming code and thus have no physical components at all—can (as "trustees," at least) be in trust relationships with HAs in a diffuse-default-trust environment. With respect to (b), the level of trust can vary from weak (or even "minimal") to strong (i.e., "more robust"), depending on whether the

(i) contact between HAs and AAs is *direct* or *indirect*;
(ii) AAs in the trust relationships are autonomous to some extent (i.e., whether the AAs are "functionally autonomous" AAs, or FAAAs).[15]

With regard to (i): If the type of contact between an HA and AA is indirect, many different kinds of AAs (including soft bots and programs, both at the individual and multi/distributed levels) are capable of being in the zone of diffuse-default trust; but in these contexts, the trust relationship between the HAs and AAs will typically be weak. Alternatively, in trust zones where HAs have direct contact with one or more AAs, the trust relation can be considerably stronger. With respect to (ii), especially where the AA is a FAAA (and where the contact between an HA and FAAA is more *direct*), the trust

[13] As such, this model can be viewed as a "contextual theory of trust" that is similar, in some ways, to contextual models of privacy such as those articulated by Moor (1997) and Nissenbaum (2004, 2010). In both cases, contexts or zones play a critical role for understanding the nature of privacy, in much the same way that these kinds of zones also play an important role in grasping a key aspect of the theory of trust defended here. For some insights into the various conceptual connections between privacy and trust, see the collection of papers included in a special issue of *Information* (Tavani and Arnold 2011), especially the articles by Buechner (2011), deVries (2011), and Durante (2011). And for some alternative models for analyzing trust and e-trust in the context of AAs and digital environments, see the papers included in special issues of journals edited by Taddeo (2010b) and Taddeo and Floridi (2011).

[14] Note that in this essay, I focus on trust relationships involving HAs and AAs in one direction only (HA—>AA). As mentioned earlier, I leave open the question of whether AAs are capable of having trust relationships with other AAs. For an interesting discussion of AA—>HA and AA<—>AA trust relationships, see Grodzinsky et al. (2011). Also, Lim et al. (2008) briefly examine some questions pertaining to the possibility of reciprocal and symmetric trust relationships between "machines" (or AAs).

[15] Elsewhere (Tavani and Buechner in press), I have introduced the concept of an FAAA. Building on the model of "autonomous artificial agent (AAA)" advanced by Floridi (2008), as well as on the notion of "autonomous system" described in the Royal Academy of Engineering (2009) report, I argue that FAAAs can be understood as AAs that are (i) rational, (ii) interactive, (iii) adaptive, and (iv) independent (i.e., in the sense that they can exhibit at least some independence from humans). Wallach and Allen (2009) use the expression "functional morality" to describe AMAs that exhibit some degree of autonomy (as well as some degree of "ethical sensitivity"). However, the authors do not comment specifically on either the degree or kind of autonomy, functional or otherwise, that an AA need needs to qualify as an AMA (This critique of their notion of functional morality is developed more fully in Tavani 2011).

relationship can be fairly strong. In the latter case, the FAAA(s) involved will also likely be transparent, or visible, to the HA(s). Some of these FAAAs might even look and sound human-like and might also have affective components (such as "artificial emotions") built into their design.[16]

At this point, it might be useful to reiterate three key points articulated in the preceding paragraphs of this section.

1. HAs can enter into trust relationships with several different kinds of AAs, simply in virtue of the nature of the default and diffuse-default zones (of trust) involved.
2. The strength of the trust relationships between HAs and AAs in these zones can range from weak/minimal to strong, depending on the kinds of AAs involved, as well as the type of interaction (direct vs. indirect) between the agents involved.
3. Trust relationships involving HAs and FAAAs can be considerably stronger than those involving HAs and non-functionally autonomous AAs, especially in contexts where the interaction between the agents involved is also direct.[17]

Although it is possible for HAs to be in trust relationships with AAs, and while there can be varying *levels of trust* (depending on factors such as the kinds of AAs and the strength of the trust relationships involved),[18] much more still needs to be said about the specific levels of trust that are possible for the various agents (HAs, AAs, and FAAAs) comprising the zones or contexts. As noted above, a principal objective of this essay is to show how the levels of trust involved can be explicated in terms of a scheme that closely parallels Moor's four levels of ethical agents.

4 Levels of Trust Corresponding to Moor's Levels of Ethical Agency

Drawing on some key distinctions in Moor's framework, described in section 2, I propose an analogous model for analyzing trust relationships involving HAs and AAs. First, however, we should recall Moor's point that computers (and computer programs) can be agents that are normative but (also) *non-moral*. For an example of this kind of agent, consider the case of an automobile that I drive back and forth to work, which has some computerized parts in the car's engine and ignition mechanism that function as AAs. The AAs comprising this vehicle are normative in nature (i.e., they have a purpose), and so far, they have performed well, enabling my car to start successfully

[16] For example, Turkle (2011) can be interpreted as suggesting that those AAs whose appearance is more human-like can "elicit" trust on the part of an HA in ways that AAs appearing less human-like would not. We consider Turkle's point in more detail in section 5.

[17] Later in this essay, I will argue that in the case of HA–FAAA trust relationships, the FAAAs are capable of "disappointing" or "letting down" the HA. Consider that if an AA is not functionally autonomous, however, it would be odd to say that the AA let down an HA; an AA that was not autonomous (in some sense) could not have behaved differently than it did (except, of course, by malfunctioning).

[18] It is perhaps important to note that I have not discussed any concerns affecting the *trustworthiness* of the agents, as distinct from the trust relation itself. (Whereas philosophers typically regard trust as a certain kind of *relation* between two agents—a trustor and a trustee—trustworthiness tends to be viewed as a *property* or *characteristic* pertaining to the trustee.) In a separate paper (Buechner et al. 2014), I consider the question whether there can also be levels or "degrees of trustworthiness" for AAs in trust relationships affecting HAs and AAs and, if so, whether they also parallel the four levels of trust articulated in the present essay.

each day. Because my car has been very dependable in starting whenever I turn the ignition key, I have come to rely on my car starting each morning. However, can I reasonably say that I "trust" my car to start tomorrow morning? While I may have an expectation that my car will start tomorrow, this expectation is not one that is necessarily normative in nature. Of course, some additional factors affecting my expectation also need to be taken into consideration—e.g., how old is the car? How old is the car's battery? Has the car been recently serviced? If this car is relatively new and has been serviced at appropriate intervals by a qualified mechanic or car dealership, I might be in a "zone of default trust" that includes several parties (i.e., "agents"). So, under certain conditions, it is conceivable that I might also have some normative expectations about my car's starting tomorrow. However, a more important question for us to consider in this context is whether my car's failing to start tomorrow will necessarily have an ethical impact.

What would the typical kind of *impact* likely be for me (and for others) if my car fails to start tomorrow morning? It certainly might cause me some inconvenience; for example, I might have to call a friend or hire a taxi to get to work, or perhaps I will need to find some alternative form of transportation. But such an inconvenience would not typically have an impact that is ethical in nature. In this sense, the AA (or system of AAs) responsible for starting the engine for my motor vehicle would clearly seem to fit Moor's notion of a non-moral, normative agent. (And, perhaps more importantly for our purposes here, this kind of AA need not figure into any of the four levels of trust that I articulate in this section.)

Next, we consider a variation of the example of my car failing to start. In that scenario, however, the AA(s) involved function in a way that would seem—i.e., could plausibly be interpreted—to have at least some ethical impact.

4.1 Level 1: Trust and Ethical-Impact Agents

Suppose that I have recently purchased a new automobile (say, for example, a top-of-the-line Lexus) that has an "intelligent lock-unlock/ignition" (ILI) system. This system is designed such that it can recognize (or "sense") me when I approach the car and then unlock the car's doors and start the engine automatically for me. As a safety feature, no one who fails to be "recognized" by the ILI system can unlock the car's doors; and even if someone breaks into the automobile, ILI will prevent the person(s) from successfully starting the engine. However, once I have entered the vehicle—and only after I have done that—I can modify the ILI system internally such that the car will be able to be unlocked and started by a garage attendant or by a mechanic who needs to service the vehicle. I feel pretty confident that with this brand new vehicle, and its ILI system, I can rely on my car to start each morning (and I can also rely on it to be secure from automobile thieves). But can I "trust" this car, or its ILI system (with its various AAs), any more than I was able to "trust" the car involved in the previous scenario? It would seem that the two scenarios are similar in that both involve one or more AAs that are normative in nature but which also seem to be non-moral because these AAs do not function in ways that typically would have any ethical impacts.

But next consider the following circumstance that suddenly arises—at approximately 3:00 AM (03:00 h.), my wife (Joanne) wakes up with severe chest pains. Fearing that she may be experiencing a heart attack, or some other life-threatening condition, she

urgently asks me to drive her to the nearby hospital. (We both believe that it may be too risky to call the 911 emergency phone number at this point and then have to wait a short period for an ambulance to arrive.) So, Joanne and I frantically rush to our car; but we are surprised to discover that the vehicle's ILI system fails to unlock and start the car. (Perhaps the ILI has never "seen" an image of me or Joanne at 3:00 AM, especially against the backdrop of the misty fog that has set in, and thus fails to "recognize" us.) Next, I realize that I cannot physically break into the car without also automatically disabling the car's ignition system. Joanne and I panic because precious time has been lost, and her condition appears to be worsening with each passing minute. We fear that if we do not soon get to the nearby hospital, Joanne's life may be at risk.

Is the kind of impact involving the AA(s) vis-à-vis my car's failure to start in this scenario still merely (normative) non-moral, or does it now fall into the ethical-impact category? It would seem that the AA(s) comprising the ILI system in my car now have an ethical impact and thus, on Moor's criteria, could indeed qualify as ethical-impact agents.

Next, consider a scenario where a computerized component in my car that has some "decision-making" capability (and perhaps even some—albeit, very limited—autonomy): the car's anti-lock brake (ABS) system. This system more clearly satisfies the criteria for a conventional AA, in one sense, because it "acts" (or is capable of acting) independently of me or, or any other human driving that vehicle, once the vehicle's engine is running and the car is on the road.

4.2 Level 2: Trust and Implicit-Ethical Agents

Suppose that my new Lexus automobile has a state-of-the-art (computerized) ABS. In my trust relationship with the Lexus/Toyota Corporation (a diffuse-default zone of trust involving many HAs and AAs), I have a normative expectation that my car will function properly in an ice/snow storm that is forecast for later today. How, if at all, is the potential trust relationship involving the AA(s) in this scenario different from the scenario where I merely relied on my car starting tomorrow or from the scenario where I depended on my car starting to drive Joanne (my wife) to the hospital in a timely manner? For one thing, my car's ABS is designed to "decide" when to engage (and then disengage), in situations involving adverse weather conditions. Because I do not interact *directly* with my ABS, I may not consciously think of it *as* an AA; but my ABS can be viewed as an AA that is designed to "prevent harm" (in the sense described in Moor's category of implicit-ethical agent, and especially in connection with his example of the auto-pilot system in an airplane). So, the level of ethical impact involving my ABS in this scenario would correspond to Moor's category of "implicit-ethical agent." And, arguably, I place a higher degree of implicit trust (or at least dependence) in the AAs comprising my ABS than in the AAs comprising my car's ILI system. Furthermore, this higher level or degree of implicit trust would seem warranted, given the potential risk/vulnerability for me.

Additionally, the *stakes* involved in this scenario would seem to be much higher than in the case where my car's failing to start delayed Joanne's getting to the hospital in a timely manner. Whereas the failure of my car's starting in that case may have put Joanne's health/life in jeopardy, my car's ABS failing to engage properly can result in a serious traffic accident causing injury or possibly even death to me (as the driver of this

vehicle), any passengers who may also be in my vehicle, and to (potentially numerous) other motorists who happen to be traveling on the same road and within a certain proximity to my car. While issues affecting stakes extend beyond the model of trust vis-à-vis ethical agency proposed in this section, I briefly return to this topic again in the closing paragraph of section 5.[19]

4.3 Level 3: Trust and Explicit-Ethical Agents

Continuing with examples involving automobiles and AAs, I next decide to purchase a (newly available, government-approved, and arguably safe) "smart car" that does not require a human driver. After more than 30 years of driving several times a year from New Hampshire (NH) to Pennsylvania (PA) to visit family members for holidays, important family events, etc., I have reached the point where I can no longer tolerate having to physically drive that (350-mile) distance; so I am very excited about my recently purchased smart car that can now transport me safely between NH and PA (as well as to and from destinations involving other long and short trips, both outside and within my local community). Initially, however, I was not totally comfortable with owning a (fully) driverless vehicle, so I decided to purchase the model that comes equipped with a "robotic chauffer" physically located in the driver's seat—i.e., a human-like (robotic) entity similar to "Johnny Cab" in the well-known (1990) movie *Total Recall*. In light of that film, I decide to name my robotic chauffer/driver "Johnny Bot."

It turns out that Johnny Bot is not only an excellent "driver," but in recent months "he" has also become a kind of "companion" to me. He knows many of my tastes and preferences—e.g., he knows that I prefer traveling on roads with either no or very limited truck traffic; that I enjoy listening to classical music radio stations in the car; and that I like stopping at Starbucks coffee shops, etc. So, on our various trips, Johnny constantly informs and updates me regarding the various road options available to our destinations, various classical music radio stations that are playing pieces that closely correspond to my mood at a particular time, and the proximity of various Starbuck's coffee shops on the roads that we travel. (Johnny has even become very "knowledge-able" about classical music, and "he" and I have some very stimulating conversations on the trips I have taken in my relatively new smart car. For example, one day Johnny asks me if I had ever thought about the reasons why Jan Sibelius did not compose an eighth symphony; I respond to Johnny that this question had not occurred to me but that his question has now made me curious.) Because Johnny appears so human-like in so many dimensions, including the wide range of facial expressions and body gestures that he can display, I find myself beginning to trust "him" in certain ways (i.e., "he" has begun to elicit trust from me[20]).

But now consider, once again, the above scenario involving the sudden onset of Joanne's (my wife's) illness at 3:00 AM and my effort to drive her to a nearby hospital. One difference now, however, is that because Johnny "knows" (i.e., at least recognizes or "senses") me, Joanne and I have no problem entering our car and having its engine

[19] For a fuller discussion of issues affecting the role of *stakes* in various kinds of trust relations, see Carr (2012).

[20] Turkle (2011) refers to this phenomenon as the "Eliza effect," in light of the way in which Joseph Weizenbaum's "Eliza" program was able to "elicit trust" on the part of some humans who interacted with that computer program.

started automatically. And because I am so distressed about Joanne's situation, I am relieved that Johnny will drive us both to the nearby hospital. It turns out, however, that there are four traffic lights between my house and the hospital; but at 3:00 in the morning, there is virtually no traffic on these lesser-traveled roads. So there should be no problem, even if all of these traffic lights are red when we encounter them (since we could safely drive through them). The traffic lights are also synchronized such that if a motorist encounters a red light at the first intersection, the three subsequent traffic lights encountered will also turn red just before the vehicle approaches them. So, as we approach the first traffic light, which has just turned red, I ask Johnny to drive through it (since there are no other vehicles around). But Johnny informs me that he cannot do this, because his job is to protect me (and any other passengers who happen to be in the car) by driving safely and that this means following the rule: "Always stop at a red light." Being a very sophisticated FAAA, however, Johnny also has some alternative ("override") rules built into his decision-making process that anticipate dilemmas involving certain kinds of scenarios—for example, one such rule is the following: "It is permissible to drive through a red light if the vehicle is in danger of being hit by (or colliding with) another vehicle" (because driving through the red light in this instance will be safer overall for those in and outside the vehicle). However, no such rule as "It is permissible to run a red light in order to reach a destination more quickly because of a health/medical emergency" has been included in Johnny's set of decision-making/override rules; so, Johnny will not drive through the traffic light while it is red. I now plead with Johnny, but he will still not drive forward until the traffic light turns green. So, finally, I try to override Johnny's decision and attempt to take control of the vehicle manually. But Johnny resists my attempts because "he" (or his circuitry) interprets my request and subsequent behavior as "irrational." So, as in the previous scenario, precious time has once again been lost in getting Joanne to the hospital.

How is this scenario different from the earlier one involving my car's ILI system failing to recognize me, thus preventing the car from unlocking and the engine starting successfully so that I could drive Joanne to the hospital? For one thing, while the AA in the earlier scenario failed to act (i.e., do its job) successfully, it did not "decide" to act in a way that could ultimately impact Joanne's well-being. Johnny, on the other hand, had some explicit decision-making capability, even if "he" did not have full autonomy. In one sense, "he" behaved as an FAAA, demonstrating functional autonomy in critical situations. And because Johnny also exhibited some human-like qualities, he had elicited some (perhaps low-level) form of trust from me. But more importantly, Johnny's actions not only disappointed me—as, for example, when I was disappointed that I could not enter and start my car at 3:00 AM (in scenario 1)—he also "let me down" (even if unintentionally). Fortunately, however, we can assume that Joanne did eventually get to the hospital in time and was then routinely discharged without incident. But in retrospect, it would seem that I may have placed a degree or level of trust in Johnny that was not warranted.

4.4 Level 4: Trust and Full-Ethical Agents

Following the incident involving Johnny Bot (in the preceding scenario), I decide to replace my (driverless/robotic) smart car with a (brand new) conventional car, which also has many computerized parts in the form of AAs, but which also enables me (and

other HAs) to have complete control of the car. In light of the preceding incident, Joanne and I both feel much more comfortable having regained total control over driving the new car we recently purchased. One downside for me, however, is that I now have to physically drive on the trips to and from PA. And, as it so happens, Joanne and I are planning to drive to PA in a few days to attend a nephew's wedding. It also turns out that I have recently undergone a medical procedure and, because of it, was strongly advised by my doctor not to drive for the next 2 weeks (and that if I absolutely had to drive, only to do so locally). Fortunately, Joanne has agreed to drive for the entire length of the upcoming trip to PA. So I feel a bit relieved.

Suppose, however, there has been a great deal of tension involving Joanne and her family members living in PA during the past week or so—partly because of issues in finalizing some details for the upcoming wedding and partly because of health issues involving her 94-year-old father (who also lives in PA). So Joanne now finds herself under considerable stress, as the date for our trip to the wedding in PA approaches. Because Joanne is experiencing increased anxiety, she consults her physician who, in turn, prescribes a very mild anxiety medication for her (in the short term). Fortunately, this medication seems to be working out well—i.e., she has no reactions or side effects—so we feel confident about her ability to drive to PA. But during the night before we are to leave for our trip, Joanne experiences severe anxiety and is unable to sleep. Realizing also that she is the one who will be driving in the morning, she begins to panic about not being able to sleep. So, she decides to double the dosage of her medication, believing that because her anxiety is now so intense, the increased dosage will not adversely affect her. In the morning when we are ready to leave for the trip, she determines that it would be best not to tell me about her decision to increase her medication the night before (because Joanne feels that she will be fine as the day goes on, and she does not want me to worry needlessly).

However, as we are riding together, I notice that she appears to be very groggy, and I also notice that her reaction time behind the driver's wheel seems slower than usual. Eventually, we are involved in a minor, "fender-bender" accident; fortunately, however, no one is seriously hurt. At this point, Joanne informs me that she had doubled her medication dosage during the previous night. I then respond by saying: But I always trusted you in these matters and felt that I could always count on your confiding in me on decisions that could have such a critical impact for both of us.

How is this scenario different from the one involving my trust relationship with Johnny Bot? Although I had begun to place my trust (at some level, at least) in Johnny, and while "he" (unintentionally) disappointed me and let me down, Johnny did not betray my trust in him. Consider that because Johnny had limited autonomy (as a kind of functionally autonomous AA), he could not freely have done other than what he did in that particular situation, given the specific software programming code built into him. But Joanne not only let me down, she betrayed my trust in her because she intentionally withheld information from me that, in the past, I had always trusted her to be forthcoming with me. Because Joanne is a full-ethical agent in the sense described by Moor (i.e., someone who has consciousness, free will, and intentionality), she was able to betray my trust in this situation in a way that Johnny Bot had not and could not have. Consider that, unlike the level of my trust relation with Johnny, the level of trust that I was capable of placing in Joanne (as well as in other HAs) was full and complete.

5 Key Variables for Determining the Levels of Trust

Each scenario in the preceding section was designed not only to illustrate the four levels of trust that correspond to the four levels of ethical agents articulated in Moor's model, but also to argue for the importance of differentiating levels of trust in the context of AAs. But how, exactly, do the four scenarios also help us to determine the specific level of trust that would apply to a particular AA (or particular kind of AA)? What these scenarios reveal is that three key variables need to be taken into consideration to determine the appropriate level of trust:

(I) *Autonomy* (involving the individual AAs);
(II) *Risk/vulnerability* (on the part of the HAs that placed their trust in AAs);
(III) *Interactions* (direct vs. indirect) between HAs and AAs.

Regarding (I), we saw that there were different levels of autonomy and decision-making abilities on the part of the various AAs included in the four scenarios. Generally, the more autonomous the AA, the stronger the HA–AA trust relationship will likely be.[21] This factor can also help to demonstrate the differences in the kinds of "weak" (and even minimal) trust relationships that are possible in diffuse-default environments (such as those involving car manufacturers that provide many non-ethical-impact and low-level-ethical-impact AAs in their automobiles) and the kinds of stronger trust relationships that are possible between HAs and FAAAs such as Johnny Bot. Finally, these cases also illustrate why a complete or "full" trust relationship is only possible between HAs, at least at this point in time (despite the ongoing development of increasingly sophisticated AAs).

Of course, one might initially assume that it would be prudent for HAs to place less trust in AAs with higher levels of autonomy. For example, the more autonomy an agent *qua* trustee has, the greater the opportunity the trustee has to violate the trustor's trust. So, on this view, it would seem that I should place less trust in Johnny Bot than I would in the ILI system (described in the scenario illustrating level-1 trust). But we must also look carefully at the nature of the "strength in the trust relationship" itself between an HA and an FAAA such as Johnny Bot (illustrated in the level-3-trust scenario), compared to a trust relationship involving an HA and a lower-level AA such as the one described in the scenario involving the ILI). Perhaps, an analogy describing a trust relationship that involves only HAs would be useful here to illustrate the correlation between stronger trust relationships and higher levels of autonomy (for the agents involved). Consider, for example, a trust relationship involving my daughter and me. We can assume that my 25-year-old (adult) daughter now has far more autonomy than she had when she was 6 years

[21] Some also suggest that the more human-like the AA appears, and the more emotion the AA seems to exhibit, the greater the amount of trust the HA will likely be willing to place in it. For example, Coeckelbergh (2010, 2012) suggests that for humans to trust AAs, future AAs will need to have some affective/emotive qualities built into them and will need to (physically) appear more human-like. Along somewhat similar lines, Turkle (2011) suggests that the more human-like an AA looks and behaves, the greater the amount of "attachment" and "trust" the HA (who interacts with such an AA) may be disposed to place in it. However, it is also worth pointing out that others believe that the "uncanny valley" hypothesis (in which robots that appear "almost," but not exactly, human-like" can repulse humans) might have the opposite effect on the amount of trust and attachment that an HA would place in the AA. Regardless of which view turns out to be correct, it would be useful to disentangle the concepts of attachment and trust in contexts involving AAs that appear human-like, because of the way in which the two concepts can become so easily convoluted in such contexts.

old; so, it would also seem that there might be a greater risk involved in my placing trust in her now (to carry out some action, X) than there was for me to do so when she was a child. But, on the other hand, I am capable of having a much *stronger trust relationship* with my (fully autonomous) adult daughter than I could have had with my 6-year-old daughter.[22] Along with that stronger trust relationship, however, comes a greater risk or vulnerability for me as a trustor. For example, my adult daughter is not only capable of letting me down or disappointing me by her actions (as she also could when she was still a not-yet-fully autonomous person), but she can now, as a fully autonomous agent, also betray the trust I placed in her (in a way that she could not have as a child).

So we can see that there is a direct correlation between (II)—the level of risk/vulnerability on the part of the trustor—and (I), the level of autonomy of the trustee. The higher the level of autonomy the trustee has, the higher the level of risk/vulnerability for the trustor; but, we have also seen that the higher the level of autonomy the trustee has, the stronger the trust relationship between the two agents (trustor and trustee) can be. With regard to (II), we also saw that the different degrees of risk/vulnerability for HAs (such as in the scenarios involving Joanne and me) in "trusting" one or more AAs can range from mere inconvenience at one end of the spectrum to potential loss of life at the other end.

Regarding (III), we saw that different levels of interactions, between direct and indirect/diffuse, also affect the amount of trust an HA would generally accord to one or more AAs; typically, the more direct the interaction, the greater the amount of trust an HA may be inclined to place in the AA(s). A detailed analysis of these three key variables would require a separate paper. However, Table 1 illustrates (in summary form) some correlations involving levels of trust and levels of ethical agency vis-à-vis the three variables (autonomy, risk/vulnerability, and interaction).[23]

An additional variable or element, briefly introduced in the preceding scenarios but not elaborated upon in this section, has to with the kinds of *stakes* involved—i.e., stakes affecting the outcomes for the HAs vis-à-vis the AAs' "decisions," actions, or failure to act. For example, there were significantly different kinds of stakes involved in the outcomes affecting the four scenarios; consider the number of additional persons whose lives were potentially at risk in the second and fourth scenarios, when compared to the first and third scenarios. However, further analysis of this variable would take us beyond the scope of the present essay.

6 Concluding Remarks

The main objective of this essay has been to show why it is useful to differentiate the various levels of trust that are possible in HA–AA trust relationships and to show how

[22] The "stronger" trust relationship that is possible in this scenario would also seem to qualify as a more "meaningful" (and more robust) trust relationship, as well. But that point needs to be developed in a separate paper, as it is beyond the scope of the present essay.

[23] Note that the levels of weak/low/minimal trust that apply to ethical-impact agents (and possibly to some implicit-ethical agents as well) should not be interpreted in a way that suggests that these kinds of agents are to be "distrusted." In other words, the absence of a possibility for a high level of trust for some kinds of AAs should not be viewed as sufficient grounds for distrusting those AAs. Rather, trust in the context of AAs should be viewed as a kind of "threshold concept" (that agents either can exhibit, albeit in varying levels or degrees, or cannot exhibit because they fail to meet the required conditions for trust relationships).

Table 1 Correlations involving levels of trust and levels of ethical agency

	Level of the trust possible between HAs and…	Level of autonomy for AAs involved	Level of risk/ vulnerability for HAs involved	Level of Interaction between HAs and AAs
Full-ethical agents	Complete	Full	Potentially high	(Typically) direct
Explicit-ethical agents	Mid to high	Functional	Potentially high	(Typically) direct
Implicit-ethical agents	Low to mid	Partial	Typically mid	Indirect/diffuse
Ethical-impact agents	Weak/low/minimal	Minimal to none	Typically low	Indirect/diffuse

that can be done in a way that closely parallels Moor's four levels of ethical agents. In doing this, I have also expanded on a model of trust, introduced and defended in an earlier work, that employs Walker's notions of default and diffuse-default zones of trust. I believe that this model is useful in helping us to articulate some of the subtler points associated with the various levels of trust affecting AAs. However, it is also apparent to this author that much more can and needs to be said about key aspects of the interplay between trust and agency/autonomy in the context of AAs. While I must leave that project for a future work, I hope that the present essay has provided a fruitful path for further analyses of these issues.

Acknowledgments An earlier version of this essay was presented at the Second International Symposium on Digital Ethics, Loyola University—Chicago (USA), October 29, 2012. I am grateful to Jeff Buechner, Lloyd Carr, and Jim Moor for their very helpful comments and suggestions on earlier drafts of this essay. I am also grateful to the anonymous *Philosophy and Technology* reviewers for their constructive criticisms and keen insights, many of which have been incorporated into the final version of this essay. Finally, I wish to thank Joanne Abate Tavani for permitting me to include the various scenarios involving "my wife" in section 4; contrary to what these hypothetical scenarios might (unintentionally) suggest, I am pleased to note that Joanne is in excellent physical *and* mental health!

References

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge: Cambridge University Press.

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

Buechner, J. (2011). Trust, privacy, and frame problems in social and business E-networks. *Information, 2*(1), 195–216.

Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: applying the 'diffuse, default model' of trust to experiments involving artificial agents. *Ethics and Information Technology, 13*(1), 39–51.

Buechner, J., Simon, J., & Tavani, H. T. (2014). Re-Thinking trust and trustworthiness in digital environments. In E. Buchanan et al. (Eds.), *Ambiguous technologies: philosophical issues, practical solutions, human nature: Proceedings of the Tenth International Conference on Computer Ethics—philosophical enquiry* (pp. 65–79). Menomonie, WI: INSEIT.

Carr, L.J. (2012). Trust: an analysis of some aspects. Available at http://www.rivier.edu/faculty/lcarr/Trust%20-%20an%20analysis%20of%20some%20aspects.pdf.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*(1), 53–60.

Decker, M., & Gutmann, M. (Eds.). (2012). *Robo-and-Information ethics: some fundamentals*. Berlin, Germany: LIT.

deVries, W. (2011). Some forms of trust. *Information, 2*(1), 1–16.
Durante, M. (2010). What is the model of trust for multi-agent systems? Whether or not E-trust applies to autonomous agents. *Knowledge, Technology and Policy, 23*, 347–366.
Durante, M. (2011). The online construction of personal identity through trust and privacy. *Information, 2*, 594–620.
Floridi, L. (2008). Foundations of information ethics. In K. E. Himma & H. T. Tavani (Eds.), *The handbook of information and computer ethics* (pp. 3–23). Hoboken, NJ: John Wiley and Sons.
Floridi, L. (2011). On the morality of artificial agents. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 184–2012). Cambridge: Cambridge University Press.
Grodzinsky, F. S., Miller, K., & Wolf, M. J. (2011). Developing artificial agents worthy of trust: 'would you buy a used car from this artificial agent?'. *Ethics and Information Technology, 13*(1), 17–27.
Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29.
Johnson, D. G. (2006). Computer systems: moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.
Lim, H. C., Stocker, R., & Larkin, H. (2008). Review of trust and machine ethics research: towards a bio-inspired computational model of ethical trust (CMET). In *Proceedings of the 3rd International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*. Hyogo, Japan, Nov. 25–27, Article No. 8.
Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Robot ethics: the ethical and social implications of robotics*. Cambridge, MA: MIT Press.
McLeod, C. (2011). Trust. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/archives/spr2011/entries/trust/.
Moor, J. H. (1985). What is computer ethics? *Metaphilosophy, 16*(4), 266–275.
Moor, J. H. (1997). Towards a theory of privacy for the information age. *Computers and Society, 27*(3), 27–32.
Moor, J. H. (2006). The nature, difficulty, and importance of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.
Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1), 119–157.
Nissenbaum, H. (2010). *Privacy in context: technology, policy, and the integrity of social life*. Palo Alto, CA: Stanford University Press.
Royal Academy of Engineering. (2009). *Autonomous systems: social, legal and ethical issue*s. London. Available at: www.raeng.org.uk/autonomoussystems.
Simon, J. (2010). The entanglement of trust and knowledge on the web. *Ethics and Information Technology, 12*(4), 343–355.
Taddeo, M. (2009). Defining trust and E-trust: old theories and new problems. *International Journal of Technology and Human Interaction, 5*(2), 23–35.
Taddeo, M. (2010a). Modeling trust in artificial agents: a first step in the analysis of E-trust. *Minds and Machines, 20*(2), 243–257.
Taddeo, M. (Ed.) (2010b). Trust in technology: a distinctive and problematic relationship. Special Issue of *Knowledge, Technology and Policy 23*(3–4).
Taddeo, M., & Floridi, L. (Eds.) (2011). The case for E-trust: a new ethical challenge. Special Issue of *Ethics and Information Technology 13*(1).
Tavani, H. T. (2011). Can we develop artificial agents capable of making good moral decisions? *Minds and Machines, 21*, 465–474.
Tavani, H. T. (2012). Ethical aspects of autonomous systems. In M. Decker & M. Gutmann (Eds.), *Robo-and-information ethics: some fundamentals* (pp. 89–122). Berlin, Germany: LIT.
Tavani, H. T. (2013). *Ethics and technology: controversies, questions, and strategies for ethical computing* (4th ed.). Hoboken, NJ: John Wiley and Sons.
Tavani, H. T., & Buechner, J. (in press). Autonomy and trust in the context of artificial agents. In M. Decker & M. Gutmann (Eds.), *Evolutionary robotics, organic computing, and adaptive ambience*. Berlin, Germany: LIT.
Tavani, H. T., & Arnold, D. (Eds.), (2011). Trust and privacy in a networked world. Special Issue of *Information 2*(4).
Turkle, S. (2011). Authenticity in the age of digital companions. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 62–78). Cambridge: Cambridge University Press.
Verrugio, G. (2006). EURON roboethics roadmap (Release 1.1). In G. Verrugio (Ed),. *EURON roboethics atelier*. Genoa, Italy. Available at http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf.
Walker, M. U. (2006). *Moral repair: reconstructing moral relations after wrongdoing*. Cambridge: Cambridge University Press.
Wallach, W., & Allen, C. (2009). *Moral machines: teaching robots right from wrong*. New York: Oxford University Press.