A model of trust between Human Agents and Artificial Agents

Abstract

The different levels of Artificial Agents (AA), based on ethics considerations, affect the interactions Human Agents (HA) have with AA, so that they can shape in a significant way the forms of trust HAs place in AAs. On one hand weak interactions (goal-oriented interactions) occur with low level AA and imply low impact trust; on the other strong interactions (relation-oriented interactions) occur with high level AA and imply high impact trust. When developing an AA, the identification of the correct interaction the AA is supposed to have with Has (and then the identification of the type of trust) is fundamental to the process of development itself.

Abbreviations

AA: Artificial Agents

HA: Human Agents

TM: Turing Machine

EA: Ethical agents

GOI: goal-oriented interaction

ROI: relation-oriented interaction

Introduction

Due to their logic malleability, machines (hardware, software, and both hardware and software) have spread within human society. This raises questions about whether and how their usage and interaction with HA, which is increasingly widespread and intensive, involves some forms of trust, how can we define these new kinds of trust, how trust affects the usage itself, which are the key factors of trust and how these factors in turn affect trust.

To answer even to some among these questions, this paper aims at presenting a new point of view about HA – AA relations, overcoming other controversial theories and providing a useful framework to developers and engineers. Every kind of relation, if correctly defined, strongly implies a specific (and useful) definition of trust.

Firstly, I will briefly discuss about AA, recalling the theories presented in (Imma 2009) and (Moor 2006), in order to define precisely the subject. Then I will present, with the support of two examples, respectively the GOI and its features and the ROI and its features, and show that, although they may appear similar, they have completely different features and involve completely different definitions of trust. In the conclusion part I will show how this model could be used in the design phase of an AA, trying to prove its utility.

On Artificial Agents

The focus is on AA, which can be defined as artificial (in the sense that they are manufactured) entities which are able to perform actions, to *act*. In particular, following the description of (Imma 2009, page 27),

consciousness is considered necessary for agency, since acts are defined as performances caused by intentional mental states, and these states are conscious. It's not my intention to determine in an exhaustive manner here whether some artificial entities can perform actions, because it's a big and highly controversial question outside the scope of this paper, but I want just to point out some considerations which I feel relevant to connotate correctly the object of my interest.

Firstly, Artificial Entities can have internal states: e.g. the well-known notion of state of a Turing Machine, which can be simulated both mechanically by a tape and digitally by a piece of software. The internal state can have some *beliefs* (which are variables encoded in the state of the entity) and *desires* (the entity can try to do something, e.g. the *try- catch* syntax in Java, which is perfectly reproducible by a TM). A TM is conscious of its state in the sense that the entity can "privately observe" its state by simply inspecting itself, so these internal states are "mental" according to Imma's definition. An artificial entity can also change its state, or *instantiate* one, due to external motivations o to internal *intentions*: again, a TM progresses due to external input, which is provided on the input tape, but also due to internal computation, which can cause the machine to stand for certain behaviour. Moreover, in a certain sense everything a TM does is pure action, since it's always directed from its state. The machine is always conscious of it, and it cannot happen "by accident", since it's encoded in the construction of that TM. If the TM is able to instantiate internal mental states, the TM is the conscious.

In consequence, there is at least a meaning of the term "consciousness" which is satisfied by a TM, so we can state that every artificial entity which can be represented by a TM is, at least in a broad sense, an AA. It's left as an open question to determine whether this sense is equivalent to the sense we use for human consciousness, i.e. to determine whether the TM internal states are comparable (in terms of degree of complexity) with the human mental states, and whether the input tape is comparable with the human senses, and whether the two notions of intentions are comparable, but this is beyond the scope of this paper, and would lead to other controversial questions (e.g., determine if a specific artificial entity can qualify as an AA, depending on its features and the context). For the same reasons, I'm not going even to try to demonstrate that an AA could be an Artificial Moral Agent.

Note that until here I still haven't demonstrated anything, since the simple definition of TM applied to what stated by (Imma 2009) has led to this weak and "rough" definition, which, though possibly acceptable, is of little interest.

More interestingly, the analysis of (Moor 2006) goes beyond these problems, introducing a new framework: the concept of *ethical agents.* He claims that every artificial entity is, first, a *normative agent,* which means that it has been designed to satisfy some requirements and so there exist some *norms* to assess its deployment. In this sense, every artificial entity is an agent. Furthermore, these norms, or standards, can be only of practical (i.e. non-moral) nature, but may be also of ethical nature. The entities which have both kinds of norms are called ethical agents, because their performances influence ethics, and should be evaluated accordingly. Moor defines four agents, besides the non-ethical ones:

- Ethical impact agents
- Implicit ethical agents
- Explicit ethical agents
- Fully ethical agents

The difference between them is the level of ethics consciousness they have embedded in their design. On one hand, ethical impact agents, at the lowest level, are AA which do not present any ethics consideration but still have an impact on ethics, under certain circumstances. On the other hand, fully ethical agents are able to reason independently about ethical issues, argument their conclusions and take the best decision.

The only known fully ethical agent, for Moor, is an HA, but this doesn't prevent their development in the future.

The valuable aspect of Moor's work is that it helps to distinguish different kinds of artificial agents and how relevant are they for ethics considerations. We can now enforce these differentiations to my "rough" and vague definition of AA to describe the possible interactions between HAs and AAs.

Section 1: Goal oriented interactions

As an example, consider the case of the mechanical tomato harvester, a device developed by researches at the University of California from the 1940s. Its goal was to automate the operation of harvesting tomatoes in the American farms, reducing the cost of growing tomatoes with respect to handpicking, which was the common habit before the introduction of this new technology. Indeed, it doesn't need anymore farmworkers to do that job and nonetheless it's faster. On the other hand, this caused some unexpected troubles in the American society: since the machine requires a huge initial investment and works well only with concentrated forms of tomato growing, and since the job of hand pickers disappeared, the new technology ends up with causing some political issues in the American rural communities. Moreover, researchers proposed new varieties of tomato which fit better the machine process, replacing the old ones and thus having an impact on the rural environment and on the food people were used to eating. All these are ethical issues, so, according to Moor's categories, the harvester can be considered as an *ethical impact* agent. In fact, on one hand it has been designed just to solve a practical issue (how to increase efficiency of harvesting) with no ethical norms or requirements at all; on the other hand, its assessment must keep in consideration all the ethical impacts the machine had on the society (whether they were predictable or not).

Consider now the interaction of the farmer which is the owner of the mechanical harvester and the machine itself: their interaction is what I call a *goal-oriented interaction* (and the harvester is a *goal-oriented machine).* Its definition is the following: a GOI interaction occurs whenever an HA takes advantage of the working of an AA, being conscious of that and having the power to interrupt the interaction at any time, in order to accomplish a goal (delegation). Interaction takes place between a HA and an AA, which is defined according to the "rough" definition given in the first section. Moreover, the AA is at most an ethical impact agent. In the example, the mechanical harvester can be considered a TM with few states to describe the different movements to perform to pick the tomatoes, some hands to interact with the environment and the intention to pick a ripe tomato and leaving a non-ripe one. The harvester can be conscious of the actions it's performing because it's performing some specific and repetitive actions. The main features of this interaction are:

1. It happens just for one specific practical goal;
2. It happens seldom: the owner buys the harvester once and uses it just for a limited period of the year. The definition of "seldom" depends here on the context, and a useful tool for assessment this point is dependency: How much of HA business is delegated to the AA? Could the HA continue to do its business (maybe with small damage) if the AA suddenly crashed? In the example, a crash of the harvester in a season other than summer would not affect in any way the farmer's activity, while a crash in summer could be tackled by reintroducing immediately hand picking. It's remarkable here that dependency in some cases could not be foreseen during the design phase of an AA, since it's influenced by the actual use HAs make of it.
3. The human knowledge about the task performed is comparable with the maximum amount of knowledge the machine can have from any kind of source. This does not prevent the machine from being able to perform the task in an incomparably better way.
4. The interaction can be indifferently physical (face to face, *f2f*) or digital (i.e., electronically mediated): since these interactions are not strong, digital ones are more frequent. In my example, the owner

could activate the harvester by a button on the machine or remotely in the case the machine is connected to Internet.

5. The interaction can be indifferently direct or indirect: since these interactions are not strong, non-direct ones are more frequent, although my example is a direct interaction one. I define *indirect interaction* an interaction when the HA is still conscious of the existence of the interaction, but it's not directly interfaced with the AA, because of an intermediary (which is not merely a communication channel, but another AA).

Trust in goal-oriented interactions

Based on the approach used, the definition of trust in this kind of situations is rational. Given an HA (the *trustor)* and an AA (the *trustee)*, and an interaction between them in order to achieve a particular goal(GOI), trust is the constitutive element of the interaction, in which the trustor evaluates on one hand the benefits of committing some actions to the trustee, and on the other hand the risks connected to the fact that the machine can fail in performing in his actions or damaging the trustor in any other way during the interaction, and chooses whether starting (or continuing) the interaction (if the benefits overcome the risks) or not starting (or stopping) the interaction. Furthermore, a positive but not sufficient (depending on individual considerations) level of trust or simply a high level of risk could lead to a form of supervision: in this scenario the trustor applies some forms of control on the work of the trustee. The cost (in terms of resources) of this control reduces the benefits of delegation, but it's then balanced by the reduction of risks, so the net value of trust remained unchanged. Trust could be also seen as a value of the expected advantages which come from the interaction to the trustor. In case of multiple candidate trustees, the assessment is performed to select the best AA to trust. A relevant point of this definition is that trust is evaluated for a single interaction, and interaction concerns a single goal: for example, the farmer evaluates the harvester for its ability to harvest tomatoes, and for nothing else. This is the only point he cares about because it's the reason for the existence of the interaction, the *goal.*

I want to stress out that this definition is quite similar to the one given by (Taddeo 2010), which states that trust is a "second order property that affects a first order relation between two agents" (pag.255). This property consists of "minimizing the trustor's effort and commitment in the achievement of the given goal" (pag.249). As in Taddeo, I define trust as related to what I called interaction (and Taddeo "relation") and not standing alone (i.e., both are goal-oriented). I disagree with Taddeo on the fact that I consider trust as *part of* the interaction in the sense that trust is always the starting point of the interaction, and it's not merely a "property" of interaction. Moreover, my analysis of the supervision is different: Taddeo states that supervision reduces the benefits of the relation, since it forces the trustor to spend additional efforts when the trustee is not enough *trustwhorty,* with no advantage ("the minimization of the trustor's effort and commitment allows the trustor to save time and the energy that it would have spent in performing the action that the trustee executes, or in supervising the trustee.", pag. 251). On the other hand, in my approach every interaction contains in itself a level of risk, which must be taken in account when considering the total amount of resources to be spent the in the interaction. Supervision reduces this amount, so it doesn't not affect in any way the net value of trust. Ease of supervision provides the HA an additional tool which can be used without any cost. Moreover, due to feature 3, it's a general property of this kind of interaction. Ease of supervision is correlated to the ease of assessment trust, which holds in every moment: the HA can decide to use the partial past results of the interaction to assess a new level of trust and deciding whether going on in the interaction. This process is still completely rational, since it's carried out by simply inspecting AA's past actions.

Some authors (e.g., Tavani 2015) have defined trust as the expectation that the AA will accomplish the given task. I claim that this definition is incomplete, because it lacks the consideration of risks: the HA decides to delegate an action to an AA with the expectation of future advantages (the accomplishment of given goals) *in spite of* the risks he is going to take, ad trust is a sum of both advantages and risks (each with its probability). Trust is then the expectation of overall advantages, given that accidents could always happen.

Considerations

One might put the emphasis of the result of the assessment instead of the assessment itself and define trust as the expectation of advantages (as I just did) from an interaction with an AA. This is still acceptable but less useful for the analysis.

This process of assessment is both rational, since it's done considering only facts and probabilities of future facts, but also subjective, since it depends on the weight the trustor gives to each objective risk and benefit. To evaluate trust, a model like the objective one (of thresholds) presented in (Taddeo 2010) could be adapted with not much effort to suit this definition. Indeed, it's still opened the possibility of the introduction of frameworks to compute trust as a function in which the trustor inserts as input his personal weights of facts, and facts are represented as number correlated to their potential effects multiplied by their probability of occurrence, which give *statistical* benefit (previsions) and *statistical* harms (risks).

One might observe that this rational evaluation isn't necessarily carried out before each interaction, even if the HA is conscious of that interaction, or it's only partially carried out. This happens for various reasons (HA doesn't care, lack of knowledge, lack of time for an in-depth analysis…). The model presented handles these cases simply by putting 0 as input parameter for that elements whose evaluations hasn't been carried out, so that they don't count in the function of trust, as if they didn't exist. Indeed, trust can be seen as the sum of infinite addends, where the HA puts 0 as coefficient of all the addends whose existence he doesn't know or doesn't care.

Given the definition and the extension of the model in the previous paragraph, it makes no sense to ask whether there is need to "trust" an AA, since trust's assessment is seen as an essential part of each HA AA interaction, and takes place even if the HA is not conscious of it.

Due to feature 3, trust can be easily assessed in any time of the interaction, so that the HA can interrupt the interaction whenever he realizes that context has changed, and trust is no more advantageous (e.g., because AA has become old). Past interaction is useful to assess new levels of trust.

Trust could be "betrayed", in the sense that an interaction turns out to be disadvantageous for the trustee, contrary to what assessed before. The damage is limited to the goal of the interaction, and besides this the HA has the opportunity to anticipate it through the supervision of the work of the AA.

Note that I've completely avoided to discuss all the differences between a *f2f* interaction and a e-interaction, although it's considered a relevant point in trust-related questions (as a reference, see (Grodzinsky and alia, 2011)), because I believe that they don't matter in my approach, while other authors put much effort on this topic, e,g. (Taddeo, 2011). I admit that this is a weakness of my model which would need a more in-depth analysis.

Section 2: Relation oriented interactions

Consider now a case which is related to the one presented in the previous section: a farmer decides to use a series of robots to fully automate his farm, from the planting process to the picking process. Although this scenario was unrealistic in the 40's, it's not so far from reality now. All the robots are guided from a boss (which is a robot too), who takes all the decisions. In developing the robot boss, the engineers must take into account a lot of ethical questions: respect the animals, do not damage the environment with some pollutant herbicides, do not waste resources like water, defend the cultivations from external agents (like insects) but do not destroy the ecosystem. Since the robot contains in its code the rules necessary to cope with these issues, it is an example of an explicit-ethical agent, according to Moor's definition. Its features comprehend not only the execution of manual tasks (the coordination of other robots), but also the ability to "reflect" on ethical issues and take a decision accordingly. This ability is anyhow limited, since it depends only on how it has been designed, so that it can't modify by itself the way it evaluates events and take decision. In other words, it "reasons" always following a preestablished pattern, the one encoded by the developers.

Although it might be similar, the interaction between the farmer and the AA is of a completely different nature: we can identify a specific goal (the management of the farm), but the delegation of the accomplishment of this goal has a notable impact on the farmer's life (it's almost his entire job) and makes now his subsistence totally dependent on the correctly working of the AA. This kind is what I call a *relation-oriented interaction* (and the robot is a *relation-oriented AA).* This interaction takes place between an HA and AA, defined according to the "rough" definition given above, which is also at least an implicit ethical agent, according to Moor's definition. The definition of this interaction is rather like the GOI: it occurs whenever an HA takes advantage of the *support* of an AA, being conscious of that and having the power to interrupt the interaction at any time. Intentionally there is no mention of the goal and it's used the word "support": I want to highlight the pervasiveness of the interaction. Therefore, the features of the interaction are:

1. It happens for one goal, which is not only practical but includes some form of (explicit or implicit) reasoning about ethics and consequences of actions.
2. It happens frequently: as above, dependency and quantity of delegation is an objective tool to assess this.
3. The AA could obtain a knowledge which is incomparably greater than the one of the HA: this makes the supervision much more difficult and leaves the control to the machine. This is a key point which will lead to a different discussion of the kind of trust involved here. In my example, the robot could have a technology to consult and read in a few seconds all the encyclopaedias of the word (that is, a simple connection to Internet and a high computing capacity), while every HA, even those experts in farming, couldn't obtain all that knowledge, since it would take centuries to them to read all the encyclopaedias.
4. The interaction can be indifferently physical (face to face, *f2f*) or digital (i.e., electronically mediated)
5. The interaction is direct or indirect: since its anyway strong, direct interactions are more typical.

It's remarkable here that the AA can have human appearance: for example, it may use a human like voice to communicate, or it can simulate human feelings, or it could look like

Trust in relation-oriented interactions

To define trust in this kind of interaction, we must firstly look if the previous definition of trust is applicable here too, considering the new features. While evaluating benefits could be considered still possible, evaluating risks causes some troubles, since there is strong delegation and supervision is very difficult to perform, so that after some time it's very difficult to assess trust again. Moreover, we must take in account the new features of the interaction. Since it's very strong, the HA feels that the AA has a great presence in his life, so the interaction introduces some form of human feeling, like empathy. Due to that, it's no more purely rational. Given a HA (the trustor), an AA(*the trustee)* and an interaction between them in order to

build a relation(ROI), trust is the constitutive element of the interaction, by which the HA evaluates *before starting the interaction* on one hand the benefits of being supported by the trustee, and on the other hand the risks connected to the fact that the machine can fail in giving support or damaging the trustor in any other way during the interaction, and decides whether starting the interaction. During the interaction, trust involves more and more feelings, which then play an increasing central role in assessing trust, such that it's impossible to set up a model for trust without the help of psychology. After a certain amount of time, benefits will regard mostly the psychology of the HA, while risks grow exponentially: trust has completely changed and now it's imponderable.

I want to stress out that in this context trust does not pertain only a specific or practical goal, but the *existence* itself of the AA: in my example, the farmer trusts the robot in the sense that he feels that the robot will not only be able to harvest the tomatoes or feed the animals, but also to cope with any unexpected issue which could threaten the crop safety. The benefits involve the feeling of security of the farmer, while risks are very high: if the robot doesn't manage to defend the crop from thieves or insects, the farmer will starve. Since the robot takes the control of all the farm, it's very hard for the farmer to supervise all its actions, and so there is no way to determine a new rational level of trust based on past interaction. Every rational assessment would be very cautious in conceding trust again (i.e., deciding to continue the interaction), but some feelings intervene to affect the new assessment in an imponderable and decisive way. Trust can then irrationally interrupt the interaction or make it last.

When interaction has begun, we could define trust as the expectation that the AA will take care of us responsibly. I want to note that there is no mention of risks, which were a valuable element in the first section. This is because trust is not merely rational here, but, as said, it includes some feelings, which are necessary to develop a such strong interaction. Without these feelings, no one would take so many risks (for example, think of an HA which decides to use a self-driving machine. After buying it, and rationally evaluate its ability, the HA doesn't control anymore how it drives, since he develops a feeling of safety, but on the other hand even if he wanted the couldn't supervise every decision taken by the car, since it's taken too fast for a human mind (and maybe the HA doesn't know anything about even how to understand the programming code of the software). Would a purely rational HA use this car every day to go to work? Probably not. Trust is now of a completely different nature.

Considerations

Unlike the previous scenario, an objective model to assess and compute trust in this scenario is hard to develop, and in any case would imply a lot of approximation to include all the psychological issues. In fact, this model is not fully rational. More precisely, trust is mostly rational at the beginning of the interaction, and then progressively looses rationality, without a clear model to describe this loss.

Nevertheless, since trust is still described as a constitutive element of the interaction, also here it makes no sense to ask whether we "need" to trust AA in this scenario, since the definition of interaction implies some form of assessment of trust, which here is not more precisely specified.

The most important feature, which differentiates in the most significant way the two scenarios, is the number 3: trust can't be easily assessed because the AA has a level of knowledge (or power of computation, in some cases) which is far beyond the abilities of the HA. Difficulty of evaluation is correlated to difficulty of supervision: to supervise an AA, the HA must hinder its working, so supervision has no more

a net cost equal to zero. In some cases, the only way to supervise is to stop completely its working: in my example, the farmer should switch off the robot and inspect all the farm to supervise or evaluate it.

The betrayal of trust occurs whenever the interaction turns out to be disadvantageous for the HA, as above: as stated, it has a heavy impact on the HA. Again, I avoid any issue about the differences between f2t interactions and e-interactions.

Conclusions

The two scenarios I have described are a starter point to investigate how to encourage trust between HA and AA in different cases. Indeed, after having elaborated a model of trust, further analysis focuses on the factors of trust, i.e. how HAs assess trust, what Has view as positive for trust and what is contrarily considered negative. Trust indeed is a key factor of HA – AA interactions: without trust, all the benefits of the technological progress would remain unexploited, because interactions would be limited. It's then of great importance that developers (and engineers in general) reason about trust and its elements while designing and deploying their products (whether hardware or software or both), as well as in theoretical discussions. However, before going in details, it's necessary to have a clear model of trust: it's possible to make considerations only when it's clear which kind of interaction and which kind of trust the AA is going to generate.

As an example, all kinds of decisions about autonomy in the design phase of an AA should not be separated from the evaluation of the interaction and the trust of the AA. Consider again the case of the harvester in GOIs: increasing its autonomy means increasing its independency from the farmer. The autonomous mechanical harvester would be able to decide itself some aspects of the task, e.g. the speed of harvesting, avoiding every kind of control of the farmer. Moreover, the autonomous harvester doesn't learn from the interaction (it doesn't try to learn from the farmer how to accomplish that task or how to improve its job, there is no form of feedback, the harvester seems to already know how to do its job perfectly). Basing on the given definition of trust, autonomy has a negative effect on the interaction, since less autonomy means more knowledge about risks and easier supervision (the HA gains more knowledge about AA and how it operates and can teach the AA to adapt to his principle). It's remarkable here that a cautious HA, in case of unknown risk, will give a bigger weight to them in the mentioned function of trust, then reducing the overall measure of trust. Less autonomy means predictability, ease of recognition and assessment, which are all positive values in this scenario. In conclusion, less autonomy means reliability.

On the other hand, in the case of a ROI, the user may prefer a more autonomous AA, to strengthen even more the relation. Consider the case of the robo farmer: if the robot showed the ability to reason and debate of a child, the relation would fall immediately. Instead, a self-confident robot which appears completely independent would easily gain all the trust. More autonomy here means greater feeling of security, and more abilities to cope with all possible issues, and thus support the HA and develop a relation. An HA would never buy a robot which asks for help or advice at midnight, simply because he would consider it useless, even if it were a great harvester. The assessment of its existence indeed would give a negative result: the robot is not able to handle the relation, it's "stupid" and therefore dangerous. In conclusion, here more autonomy means more reliability.

I have just showed some considerations about a relevant topic, autonomy, which simply derive from the observation of the context and the enforcement of the model presented in this paper. These considerations can be elaborated during the design phase of every AA and improve the design itself. They could clash with other views of autonomy which lack any consideration about trust and interaction (e.g., Alterman 2000, in the context of artificial intelligence). The fact that the former ones do not only take into account trust, but

they are directly derived from the trust and interaction model gives them more strength and precision. Similarly, this model could be applied to support other relevant decisions of the design phase, like the ones concerning physical design of the AA.

References

To get more details about the mechanical tomatoes' harvester, see Wayne D. Rasmussen (1968), "Advances in American Agriculture: The Mechanical Tomato Harvester as a Case Study," *Technology and Culture*, 9, 531-543.

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.

Alterman, R. (2000). 'Rethinking autonomy'. *Minds and Machines*, 10(1), 15–30.

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.

Moor, J. H. (2006). 'The nature, difficulty, and importance of machine ethics'. *IEEE Intelligent Systems*, 21(4), 18-21.

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

Tavani, H.T. (2015). 'Levels of Trust in the Context of Machine Ethics'. *Philosophy & Technology, 28(1)*, 75–90.