Morality and Social Trust in Autonomous Robots

July 16, 2017,
MIT, Cambridge, MA

Belief, Judgment, Transparency, Trust: Reasoning about Potential Pitfalls in Interacting with Artificial Autonomous Entities

Joachim Iden, TUV Rheinland Japan, Ltd.

1. Formalizing the Notion of Trust

Definition (general): Trust of X in Y: $(X \Rightarrow Y)$ as a condition to achieve the goal G: $(X \Rightarrow Y)|_G$ is the attitude to perform action A (or abstain from performing another action B) due to the belief (subjective quantifiable likelihood) that an objective analysis of the characteristics $\chi(Y)$ would lead to the result that with respect to a suitable performance indicator π,

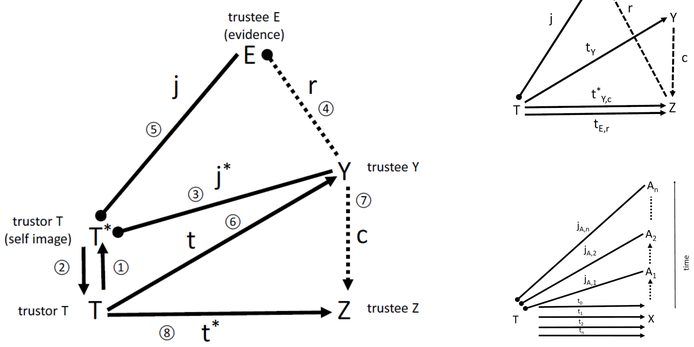
$\pi[(X \Rightarrow Y)|_G] \geq \pi[(X \Rightarrow Y_{typical})|_G]$

OR

$\pi[(X \Rightarrow Y)\,|_G] \geq \pi[(X \not\Rightarrow Y)\,|_{G^*}]$

Definition (applied): Trust is the mental attitude to forego further measures of risk mitigation due to the belief that reliance can be made on the capabilities and disposition of an agent or entity, such that if a detailed risk analysis π was performed, while taking into account all relevant characteristics χ of the agent or entity, the result of that risk analysis would show that the residual risk is of the same order of magnitude either a) as that for a comparable activity to achieve a similar goal G^* without relying on any agent or entity or b) of the same activity, achieving the same goal G while relying on another agent or entity which is well trusted by the relevant community.

2. The Trust Graph



Ascertaining self-trust (1,2) the trustor T performs an indirect judgment (3) of Y due to the belief $\beta 1$ that a relationship exists between Y and E (4), such that E is representative of Y (to degree $\gamma 1$) regarding its essential characteristics, but E is judged more easily (5). Based on that belief, T places trust in Y (6). If a relationship exists between Y and a second trustee Z such that Z depends on Y regarding its essential characteristics (7) (to degree $\gamma 2$), T places trust also in Z by association (8), again this trust may be weakened due to a less than complete belief $\beta 2$ that (Y,Z,c) holds. Self-trust is expressed in the way that the trustor's self-image formally becomes a trustee, reciprocating the trust. Trust will be misplaced ("overtrust" or "undertrust"), if the belief about the relationships between the trustees is incorrect.

Multiple Components of Trust

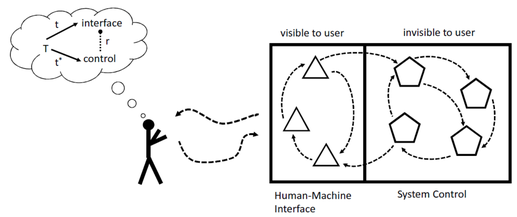
Belief in a number perceived relationships may give rise to multiple individual components of trust, either corroborating or contradicting each other.

Iterative accumulation of trust

Whenever the trustee generates a response in accordance with the expectations of the trustor, trust is corroborated. In some trustor-trustee scenarios there are sequences of discrete "reliance-compliance" interactions, where the fact of an expected response becomes evidence for the reliability of the trustee and thereby a nonagent trustee in its own right.

3. The Role of Transparency

Trust can be conferred from one trustee to another. The two trustees in question can thereby be parts or aspects of the same system. When humans interact with a technological system they do so through a human-machine interface which affords them access to display and control elements. A responsive and well designed HMI supports trust in the system. Users then often implicitly and unconsciously assume that the characteristics of the interface are representative or indicative of the whole system, including its main control subsystem.



•The questions are then whether this belief in the existence of such an indicative relationship r (interface, control, r) is justified in a specific case or even justifiable in principle.

•In social situations this leaves the trustor open to deception and there is no reason to assume something similar will not apply to interactions with artificial autonomous entities.

•The other side of this issue is whether and how transparency could be achieved for the trustor when interacting with the system. Transparency is understood here as the possibility to anticipate imminent actions by the autonomous system based on previous experience and current interaction.

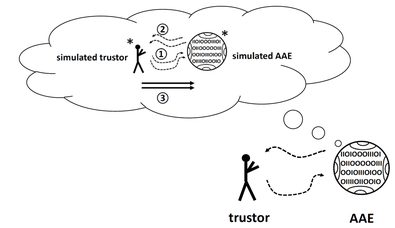
4. Morality, Empathy and Simulation

When using a product of any kind, users want to be able to trust its fitness for purpose and safety of use. When interacting with an artificial autonomous entity, we can ask what essential characteristics of the AAE may guide it to exhibit a behavior that for its trustor reflects the properties of dutifulness and absence of harmful actions. 'Absence of harm' and 'safety' in general will not only refer to protection from bodily injuries, but also to absence of psychological stress. In short, we expect the AAE to be dutiful to its purpose, refraining from activities harmful to the trustor and interacting with the trustor in a way that will be perceived as dignity and respect by the human. The latter includes also the honouring of privacy.

We can therefore say that, for the purposes of the present discussion, we are not primarily concerned with the question of moral or legal responsibility of the AAE itself, but with whether its actions are perceived as beneficial or detrimental by the involved humans, taking into account consideration of the level of difficulty of potentially required conflict resolution when more than one human party is directly or indirectly affected by the consequences of an AAE's actions. Of concern is therefore the issue of human moral patiency with respect to the actions of the AAE which we characterize as a *quasi moral agent* in the sense that we judge its actions to be moral when comparable actions by a human would be judged as moral. In other words, we derive the moral status from the *phenomenology* of its behavior, not from a supposed or constructed notion of agency inherent to the AAE.

As a model scenario for discussing the implications of implementing mechanisms allowing the AAE to operate without violating the trustor's moral integrity, we suggest to build on the concept of reciprocality and apply it to the relationship between a simulation of the trustor and the trustee implemented within the trustee AAE itself

This is expected to be comparable to a simulation of *cognitive empathy*, if the rewards for the trustee can be meaningfully defined as a function of the simulated transactions, specifically such that the trust placed by the simulated trustor in the simulated AAE ("virtual trust") serves as the AAE's reward channel. It will then also be necessary to encode the trustor's simulation such that it embodies a (purported) willingness to adhere to the ethical set of rules. The internal decision making on part of the AAE will have to be compared by it with the actual transactions involving the real trustor. At this point we could start speaking metaphorically of the (quasi) trust, the AAE places in the trustor, such that the latter becomes the AAE's own (quasi) trustee.



A) start with an initial simulation and maximize the virtual trust accumulation
B) compare the actual transactions with the virtual ones
C) calibrate the simulation against the experience
D) increase its own trust in the trustor if the experience meets its expectations from the simulation
Repeat from B)

- it may be difficult to achieve a sufficiently adequate simulation with respect to all relevant aspects
- the question of how to resolve potential conflicts when more than one trustor or more than one trustee agent are involved
- the implementation of an idealized version of the trustor ("willingness to adhere to ethical rules" in the simulation may not reflect the real trustor's actual attitude)
- in the mentioned simulation of transactions with the human trustor, repeated calibrations may not lead to a stable increase of trust on either side, trapping both human and AAE in a distrustful and potentially calamitous situation

5. Remarks and Outlook

- **AAE: A**rtificial **A**utonomous **E**ntity

- **trust** is based on beliefs about relationships ("**representativeness**" and "**control**") between entities, which may become trustees due to the trustor's own judgment and additionally held beliefs
- in real situations trustors may act irrationally, with fear, perception of convenience and addiction to a service or product affecting judgment

- further explore categories of **transparency**: a) transparency as *traceability* in design and testing, b) **interactive transparency**, c) **state-space transparency** (w.r.t. online monitoring), d) **forensic transparency** ('black boxes'), e) **'psycho-moral transparency'** as not being deceptive about the nature of the AAE, which is not a sentient being

- AAEs are considered ***quasi moral agents***, judged solely by their effect on human users

- **Future work:** expand on model in section 4: a) to study establishment and loss of trust, b) develop a detailed model of cognitive empathy

Email: joachim.iden@tuv.com