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

Joachim Iden, email: joachim.iden@tuv.com, Paper-ID [1]

***Abstract*—We present a taxonomy and formalization related to notions describing trust-based relationships in order to reason about potential vulnerabilities. We discuss the relationship between trust and the concepts of belief, judgment and transparency. Finally we present an outline of the application of some of the ideas to the problem of implementation of morality-ensuring mechanisms in artificial autonomous entities.**

Keywords: autonomous systems, trust, formal reasoning

I. THE ROLE OF TRUST

Trust can be placed in a wide and diverse range of entities. Other human beings, animals, organizations, material items or ideas. The notion of trust has emerged as a relevant topic [1], [2] when discussing human interaction with artificial autonomous entities (AAEs).

II. A TAXONOMY OF TRUST

Assuming an intuitive understanding of trust, before formalizing it in section III, we first look into the entities which will be brought into relation due to the introduction of trust. Commonly used are the terms trustor and trustee which we also adopt here while explicitly including the possibility of non-agent entities as trustees.

X: trustor, Y, Z: trustees
1st degree trust (direct trust, pre-established trust)
$X \Longrightarrow Y$ ("X trusts Y")
Example: 1) Peter trusts Paul, 2) Peter trusts manufacturer Y

2nd degree trust (associated trust)
$X \Longrightarrow Y$ and $Y \vdash Z \Rightarrow X \Longrightarrow Z$
where $\vdash$ denotes that Z depends on Y in its characteristics e.g. due to manufacture, monitoring or control
The relationship "Z depends on Y" confers trust in Z, if trust in Y is already a given.
Example: Peter trusts manufacturer Y, manufacturer Y manufactures product Z, Peter trusts product Z (that it is good for use)

0th degree trust (reflexive trust)
$X \Longrightarrow X$ ("X trusts itself")

III. 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

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

An applied example of this definition for a situation where the trustor relies on the trustee in order to achieve a goal G involving a certain risk may read as follows.

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.

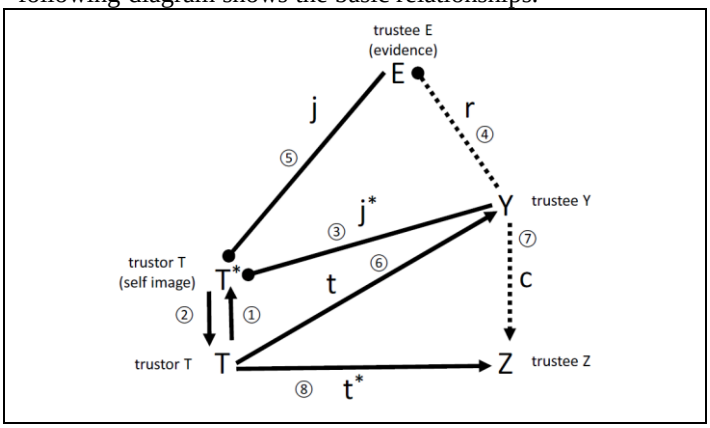
Trust may not be absolute but of a certain degree τ. In the context of risk-incurring activities relying on the trustee, the trustor will then make a comparison with activities F of a higher residual risk which do not rely on the trustee such that
$(1\text{-}\tau)\ \pi_{risk}[G(X \Rightarrow Y)] \leq \pi_{risk}[F(X \not\Rightarrow Y)]$
while $\pi_{risk}[F(X \not\Rightarrow Y)]$ is still below the level of acceptable risk for the trustor.

IV. TRUST AND JUDGMENT

A. Basic relationships of trust and judgment

Trust can be conferred from one trustee Y to another trustee Z if the trustor believes that a certain relationship exists between the two. This relationship (Y,Z,c) "control" can be described as "Z depends on Y regarding its essential relevant characteristics".

Similarly, indirect judgments are possible, given the belief in another relationship between trustees Y1 and Y2. (Y2, Y1, r) "representativeness" can be read as "Y2 is representative of Y1 regarding its essential relevant characteristics". The following diagram shows the basic relationships.



Ascertaining self-trust (1,2) the trustor T performs an indirect judgment (3) of Y due to the belief β1 that a relationship exists between Y and E (4), such that E is representative of Y (to degree γ1) 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 γ2), T places trust also in Z by association (8), again this trust may be weakened due to to a less than complete belief β2 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.

B. 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.

Let Z denote thre trustee and E_1 and E_2 two independent instances of evidence and r_1 and r_2 the respective perceived representativeness. This gives rise to two components of trust

$$(E_1,Z,r_1) \rightarrow t_{E1,r1} \; ; \; (E_2,Z,r_2) \rightarrow t_{E2,r2}$$

these two components combine to yield a resulting trust t_{res}

$$(E_1,Z,r_1) \wedge (E_2,Z,r_2) \rightarrow t_{res} = f(t_{E1,r1}, t_{E2,r2})$$

When determining the properties of the function f, we now face the choice of how to interpret the values on the trust scale. It is possible for certain scenarios to assume that all values of trust between 0 and 1 are reflecting a positive disposition of the trustor towards the trustee, no matter how small. In such a case, all sufficiently independently acquired trust components will contribute to and increase the overall trust. The contributions would then accumulate according to a subadditive function $f(t_1,\ldots,t_n)$, bounded by 1.

$$1 \geq f(t_1,\ldots,t_n) \geq \max \{ t_1,\ldots,t_n \}$$

If, on the other hand, we understand trust values > 0.5 as a positive attitude, while value < 0.5 express explicit distrust, some kind of averaging (arithmetic or geometric) must be used.

C. 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.

If a demand on a functionality by the trustor T on the trustee X, leads to the response R ($T \curvearrowright X$: R), and repeated demands lead to repeated responses

$$T \curvearrowright_n X: \{R_1, \ldots, R_n\}$$

With an assumed representativeness r for the responses of the general properties of X we can define the resulting trust after n demands $t_{res,n}$

$$t_{res,n} = f(t_1, \ldots, t_n) \text{ with } t_k = (A_k, X, r)$$

V. TRANSPARENCY, DECEPTION, RECIPROCALITY

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 question is 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.

This situation is similar to social interactions like smalltalk in humans and grooming in some social animals – trust is established on one level of interaction and the trustor then often uncritically makes the leap of faith and extends the trust by association to other levels of interaction.

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.

One can in principle imagine to equip the control system with a monitoring device which documents the internal workings of the control system and provides additional information to the trustor. The questions are then, the achievable granularity of the monitoring and whether the monitor will be effective for example in case of self-modifying algorithms without defeating their purpose.

The fact that judgments about AAEs are made indirectly via evidence such as reputation, certification or responsiveness, which are representative of the whole system only to a degree, means there are unknown and to any trustor unknowable areas of the system's state space. These unknowable areas may not only be concentrated in "one dark corner" of the state space, but densely distributed throughout it, yet undetectable due to

lack of fine graining of any practically implementable monitoring device.

In section III, reference had been made to the "disposition" of the trustee. Another way of expressing this is to consider the "morality" of the agent with respect to decisions the trustor relies upon. A number of attempts to ensure morality are discussed in the literature [3], these include approaches to constrain unethical behavior, e.g. ethical governors [4] and implementation of inherent values [1]. Here, we consider reciprocality as one mechanism that may contribute to the establishment of morality, taking hints from models of human socialization. The idea is here that there is a subset of scenarios where moral decisions of the trustee can at least be encouraged by reciprocality, viz. when the trustee values the reliance, the trustor places on it or when the trustor can genuinely offer types of interaction, valued by the trustee which it is not able to experience otherwise.

VI. 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 characterise 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.

The AAE would proceed according to the following steps

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 (quasi) trust in the trustor if the experience meets its expectations from the simulation

Repeat from B)

Problems that can be anticipated include

- 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 (its assumed "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

VII. Summary and conclusions

The main features of the presented approach are

- trustees are not limited to agents, but include both material and immaterial entities, physical objects, facts, ideas and actions
- trust is based on beliefs about relationsships ("representativeness" and "control") between entities, which may become trustees due to the trustor's own judgment and additionally held beliefs
- the relationships themselves are usually not absolute, but hold only to a degree
- trust can be accumulated iteratively through transaction sequences
- aspects or parts of a trustee may be formalized as trustees in their own right; the relationships between the parts and the whole can then influence trust in the overall system if the parts or aspects acquire trust
- the beliefs may overestimate or underestimate these degrees, resulting in "overtrust" or "undertrust"

References

[1] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016 http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[2] M. Lahijanian, M. Kwiatkowska "Social Trust: A Major Challenge for the Future of Autonomous Systems", The 2016 AAAI Fall Symposium Series: Cross-Disciplinary Challenges for Autonomous Systems, Technical report FS-16-03

[3] V. Charasi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A.F.T. Winfield, R. Yampolski: Towards Moral Autonomous Systems, arXiv:1703.04741v2 [cs.AI], 2017

[4] R. Arkin, P. Ulam, A. Wagner: "Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception, Proc. IEEE, 100 (3) (2012), pp. 571-589