Elements of a Model of Trust in Technology

Joachim Iden (joachim.iden@tuv.com)

***Abstract*—Trust is discussed in the context of other factors influencing the decision to utilize a technology and the overt and covert costs, risks and side effects incurred by that decision. We outline possible steps towards the quantification of trust in artificial autonomous systems and discuss some implications regarding the design and verification of such systems.**

Keywords: technological systems, trust, formal reasoning

I. The Relevance of Trust

Trust is neither a necessary nor a sufficient condition for the decision to rely on a specific technology, but must be seen in relation with other influencing factors and must be contrasted with the notion of *trustworthiness* with which it may be confused. In the following we understand trustworthiness as an objective characteristic reflecting the corresponding levels of *reliability*, *safety*, *security*, *transparency* and *fairness*. Trust, on the other hand, is an attitude based on subjective impression, individual judgment and experience, representing a user's ease of mind in conducting or submitting to an activity. A high level of trust may be considered appropriate when it coincides with a high level of trustworthiness, but trustworthiness does not automatically give rise to trust (resulting in 'undertrust') and trust may be found to be higher than justified when one evaluates the trustworthiness ('overtrust'). Furthermore, trust is by far the only factor affecting the decision to use a device, system or service. Other relevant factors include material availability, psychological dependency, social constraints and personally held beliefs. Further complicating the situation is, that some of these other factors may not only override the relevance of trust, but also modify the level of trust itself. In the following, we understand *rational trust* as the willingness to rely on a certain technology, being aware and accepting of the pertaining risks, costs and side effects while being (largely) free from the effects of societal or psychological constraints or amplification. We recognize that this is an idealization and a comprehensive model of trust will have to take into account and formalize the description of such effects.

II. Framework Concepts of Trust and Decision

When considering to trust a technological means (device, system or service) with respect to the capability of being helpful in achieving a goal it is in general also necessary to have a notion of what failure to achieve the goal implies and what additional factors are involved. The level of awareness of these factors may show high interpersonal variation despite the fact that in many societies relatively recent attempts have been made to highlight the impact of technology on the environment and other societies with which the users of technology may have little contact in their daily lives. Achieving a goal g comes with a direct cost c (typically a monetary value, but it could also mean permitting access to personal data among other possibilities) and inevitable side effects (e.g. environmental impact of manufacturing, operating and disposing of a technological device). Additionally, there may be hidden costs c* and hidden side effects s*. Trusting a means m implies the trustor x holds the belief, that the likelihood of there being hidden costs or side effects and the corresponding risks are sufficiently low. Finally, there are the risks of operating or using a technological artifact for a specific goal which are inherent to the technology and the mode of use. These risks, if they manifest themselves during the intended use of the device or system, result in additional costs. For a trustor x, having a goal g and a means m to achieve that goal, we summarize the framework notions of a model of *trust and decision* below. These adapt, modify and extend the notions originally introduced in the literature[1].

x trusts the means m regarding the achievement of goal g based on the beliefs in
(a) the capability of m (based on its specifications)
(b) the availability of the required functionality (based on its selectable modes of operation)
(c) the transparency of m (x believes there are no hidden costs or side effects)
(d) the fairness of m (x believes there are e.g. no algorithmic biases involved)
(e) goal fulfillment without harm (x believes the likelihood of achieving g through m is high, the risk incurred by not achieving g through m while attempting to do so is low and the risk to others is low (or x chooses to ignore these risks!)
x then decides to use m IF
(f) the level of trust T is above a determined level Tmin
AND
(g) a subjective appraisal of the total value of goal achievement, costs, risks and side effects leads to a value V greater than a determined level Vmin

III. Core Goals and their Complements

A goal g comprises the *core goal* g* (which the user ideally would like to achieve without any costs or side effects) and its *complements* with respect to its achievement, while utilizing the means m:

$$g = (g^*, g^*|_m) = (g^*, \text{costs}|_m , \text{side effects}|_m)$$

[1] R. Falone, C. Castelfranchi, "Social trust: A cognitive approach.", Trust and deception in virtual societies, Springer, 55-90

We call these the *complements of g* with respect to means m*, because g* cannot be achieved without incurring them other than by changing the means (or possibly the mode) of achievement.

With respect to risks we distinguish

- 1st person risks (risks to the trustor)
- 2nd person risks (risks to persons intentionally participating in the trustor's activity)
- 3rd person risks (risks to persons randomly encountered and not intentionally involved)
- nth person risks (risks to persons in other locations, who are usually not encountered by and whose existence may even be unknown to the trustor)
- environmental risks

IV. The Utility of Actions

An initial level of trust in a new technology can be obtained based on reputation, certification or observation of others using the technology. After that, personal experience plays an important, maybe the dominant role, while still being able to be overruled by news about accidents, research reports about design flaws, shifts in societal perception and more.

Interacting with an artificial autonomous system involves performing a series of transactions with that system, which in a sense, amounts to a form of communication. We can contrast this with the subject of classical communication theory, which mathematically describes the goal to replicate a message from a sender at the location of a receiver. Different from that, the goal of an autonomous system is to perform an action which is expected and has utility for a user. The difference between these two scenarios is that Shannon's communication theory can be considered context-free (and therefore of a *syntactic* nature) while the autonomous system's actions have a purpose (viz. fulfilling the expectations of its user) and therefore can be characterised by the term *pragmatic*. We can relate an individual action to a) the expectation by the user regarding that action, b) an intrinsic utility of the action and c) the attitude of the user. Expectation affects utility through the level of correspondence between action and expectation: the closer the match, the higher the value. If there is a complete mismatch between the two, the utility value assumes a negative sign. The intrinsic utility component is independent of this and expresses the fact that actions which pose contingent (unrelated to goal achievement) objective hazards to the user can never have a positive value. As for the attitude, we consider three states – accepting, rejecting and detached. In the accepting state, the overall resulting utility of an action has the same sign as the result determined by the expectation and intrinsic utility. In the rejecting state, utility is always negative. In the detached state, the utility value is multiplied with the imaginary unit i, expressing a situation, where the user's expectations have been frustrated to the degree that only adversarial actions are expected – any "benevolent" action by the autonomous system will then likely be ignored and its (potential) utility, while extant, becoming inconsequential.)

Symbolically, we write:

$$\text{utility}_{exp} = \langle \text{Expectation} | \odot | \text{Action} \rangle$$

$$\text{utility}_{intrinsic} = \rho;\ \rho \in \mathbb{R}$$

$$\text{attitude: sgn A; sgn} \in \{+, -, i\};\ A \in \mathbb{R}_+$$

where '<Expectation|' and '|Action>' stand for mathematical representations of expectation and action, to be combined by a suitable operation '$\odot$'.

$$|\text{utility}_{res}| = |\text{sgn A}| \times |\text{utility}_{intrinsic}| \times |\text{utility}_{exp}|$$

Whenever any of these three components has a negative sign, the sign of the resulting utility value is negative.

V. The Dynamics of Trust

Trust at a given step n in a sequence of transactions may depend on a) the value v_n of an action at that instance, b) the level of trust at the previous step T_{n-1}, and c) further characteristics of the overall time series of trust levels $\chi_{hist}(T_1,\ldots,T_{n-1})$, where χ_{hist} is an operator expressing specific properties of the time series, e.g. its variability. These dependencies need to be determined by empirical studies.

The trustor's possible actions (requests with respect to the autonomous system) and state of mind can then be modeled as a labelled state transition system (coupled to a model of the autonomous system), where trust appears as a trace history-dependent label and state transitions depend both on the value of the response by the autonomous system and the level of trust. While the actual value of trust has to be calculated at each step, for the state labelling it is only important whether that value is above or below critical threshold levels.

VI. Implications for Design and Verification

Systems do fail with a nonvanishing probability, no matter how well designed or verified they are. What one wants to avoid are systems which fail at a high level of user trust. Decreased levels of trust lead to increased caution on part of the user who may then be able to mitigate hazards which manifest themselves during system failure. This relates to aspects of transparency: a system which suddenly fails after successful longterm use, fails at a high level of experientially acquired trust and can be called maximally opaque with respect to its possible failure. Systems where diminished performance preceeds a critical failure allow the user to reduce the trust level and exercise caution during use. These can be called *'systems with portents'*. In many cases, the portents are not directly noticeable by the user and regular technical inspection or online monitoring are required to reveal them (*'systems with warning states'*). The problem is here that these measures are often insufficient to discourage the use of the system and further measures must be implemented by design (e.g. automatic shutdown and restart interlock.) From a verification perspective one wants to assure that

EITHER

the probability of transiting to a critical state at a high level of user trust is low

OR

the level of user trust is sufficiently low for all transitions leading to a critical state.