Elements of a Model of Trust in Technology

Robots, Morality, and Trust through the Verification Lens

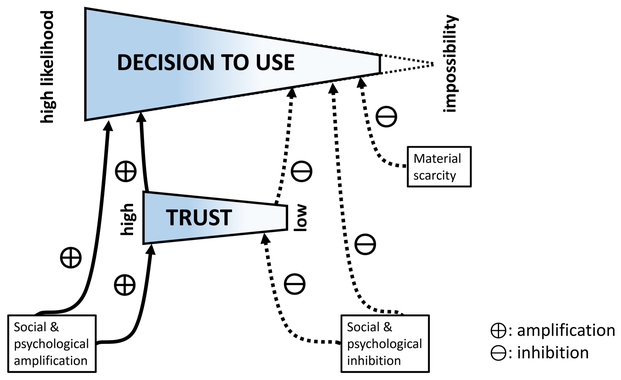
July 19, 2018
CAV-FLoC 2018 Workshop, Oxford, U.K.

Joachim Iden, TÜV Rheinland Japan, Ltd.

1. 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.

likelihood of use	social dependency	psychological dependency	material dependency
very low	ostracization	phobia	material scarcity
	discouragement	aversion	more superior than inferior alternatives
	encouragement	affinity	more inferior than superior alternatives
very high	compulsion	addiction	lack of alternatives



2. Framework Concepts of Trust and Decision

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

3. 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* $\mathbf{g^*}|_\mathbf{m}$ with respect to its achievement, while utilizing the means **m**:

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

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.

Side effects include risks.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 ri**sks (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**

goal g = (core goal g*, cost c, side effects s)

subjective allocation of value: VAL

$$VAL(<g|m>) = VAL(<g^*|m>) - VAL(<c|m>) - VAL(<s|m>)$$

objective risk determination: RISK

$$RISK(<g|m>) = RISK(<g^*|m>|x1) \wedge RISK(<g^*|m>|x2) \wedge RISK(<g^*|m>|x3) \wedge RISK(<g^*|m>|xn)$$

Risk to 2nd person

Risk to 1st person

$$VAL_{total} = VAL(<g|m>) - VAL(RISK(<g|m>))$$

4. The Utility of Actions

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 contributes to 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:

$utility_{exp} = <Expectation| \odot |Action>$

$utility_{intrinsic} = \rho;\ \rho \in \mathbb{R}$

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

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

$|utility_{res}| = |sgn\ A|\ x\ |utility_{intrinsic}|\ x\ |utility_{exp}|$

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

5. 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.

```
determine initial value of trust T0;
determine initial set of expectations e0;
while use do
    Tx(ai) ;                          /* request action ai */
    Rx(ai) ;                          /* receive action response*/
    ui = <ei|⊙|ai>;                   /* determine utility ui */
    Ti = f(Ti-1, ui, χhist(T1,...,Tn-1)) ;   /* calculate trust T */
    σ(anext): anext ∈ {a1,...,an} ;   /* determine next action request */
    enext ⟵ (anext);                  /* determine next expectation */
    if Ti < Tl ∨ Ti > Th then         /* if trust out of range */
        if ui < 0 ∨ Ti < Tcritical then
            discontinue use;
        end
        e -> e* ;                     /* adjust expectation */
    end
end
```

6. 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.

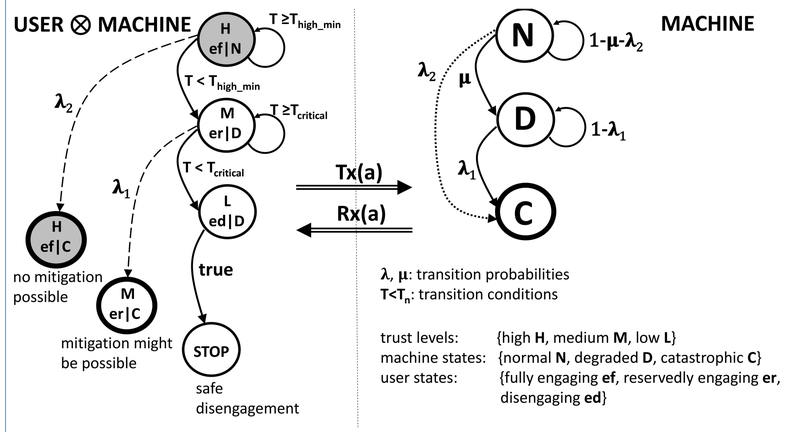
From a verification perspective one wants to assure that

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.



Email: joachim.iden@tuv.com