Trust: an analysis of some aspects [1]

Lloyd J. Carr
Dept. of Philosophy
Rivier University
lcarr@rivier.edu
Website: http://www.rivier.edu/faculty/lcarr

Trust is clearly a complex state-of-affairs. This is an attempt to distinguish some apparently significant variables involved in trust, and to suggest relationships among them. I am not here attempting to define trust, i.e. to analyze the concept; for my present purposes I accept the Tavani-Buechner definition of trust as a (voluntary) disposition satisfying certain normative expectation conditions (Tavani and Buechner, 2012b, section 3).[2] My aim is rather to explore how trust might vary as some key related features in the trust-complex vary.

The case of trust I will take as standard for my present purpose is someone (a person) or something (perhaps a collection of people) trusting someone (or some entity: perhaps an institution such as an educational or a legal system, or a business such as a hospital or drug provider, or a technology such as email or an information e-source) to do something or to be something. Here are some typical examples:

- Jill trusts her husband Jack to be faithful when he goes on business trips (= not to have an extra-marital affair)
- Jack is being sued and trusts the legal process to be fair (say, to find the law suit is meritless, if it is)
- Jill trusts her computer security program to protect her data from potential hacking
- Jack, a chef, trusts his sense of smell to alert him to food that is no longer fresh
- EU member nations trust the EU Central Bank to rescue its banks in case of a major bank failure, in the absence of a formal agreement to do so
- Drivers trust each other to drive carefully and responsibly, in the absence of continuous police presence
- Jill, who is blind, trusts her seeing-eye dog to navigate novel pedestrian obstacles

In each example the form is: someone or something (the one trusting, agent S) trusts someone or something (the trusted/trustee, agent T) to do (or to refrain from doing) something (action A). In short: *S trusts T to A* (where S ranges over agents capable of trusting, T ranges over agents able to A, and A ranges over actions in different domains).[3] Two points should be mentioned: first, while we all recognize this form of trust it isn't the only form that trust can take (see deVries 2011, and McLeod 2011). So, the analysis attempted here is restricted to a specific form of trust (a widely recognized one), and doesn't aspire to be universal. Second, both the one trusting (S) as well as the one trusted (T) are taken to be *agents*, though not necessarily moral agents, in the sense that each can effect changes by their own initiative, e.g. make a decision to trust or revoke trust, or independently (if not intentionally) perform or fail to perform an action (more on this below). This serves to rule out attributing the kind of trust I will be focusing on to non-agents, e.g. new-born infants "trusting" their parents to provide care, (some) pets "trusting" their owners to treat them well, house-plants "trusting" their owners to water them, or a dedicated robot "trusting" its user to keep it in good repair. As the above bulleted examples indicate, *agent* includes along with individual humans, groups of humans, animals, and even non-human systems. (In the example of chef Jack trusting his sense of smell, the trusted can be understood as an agent in so far as it is Jack himself exercising one of his faculties over which he has a degree of (self) control, yet is still a system that operates with a degree of independence.)

In distinguishing aspects of the trust complex and relationships among them, my main focus will be on the question of rationality. When S trusts T to A, under what conditions is S's trust rational? When would it fail to be rational or even count as irrational? In so far as trust is a choice, this question falls under the broad area of individual rational choice. My focus is motivated by a common observation: people, it seems, intuitively evaluate some cases of trust as bad judgment. In retrospect S might discover that she should not have trusted T to A, that it was foolish to have done so (I take such judgments to be saying that the trust was not rational, a bad decision). Other cases of trust are evaluated as right. In retrospect S might judge that it was a good thing to have trusted T to A, that it was the reasonable thing to have done (I take such judgments to be saying that the trust was rational, a good decision). Assuming that there *are* evaluations of trust along such lines, and given that my interpretation of such judgments is correct, it would seem that criteria are being applied, that there are norms or standards for evaluating when trust as a decision is rational and when it is not. An analysis of the kind of trust focused on here should help point the way to some of the principles under which trust is rational. I note that: (i) it is also possible to explore the rationality of the trusted's accepting the trust placed in her, and the rationality of being trustworthy (it might well be rational to reject being trusted, or to be untrustworthy, in some trust complexes), and the rationality of A'ing (A might not be a rational action in some trust complexes), but my focus here will be on rational trust and not on the rationality of other components in the trust complex; (ii) by viewing trust as a rational choice, I leave aside the kind of trust, or its equivalent, that evolutionary game theory investigates as equilibria in potentially cooperative evolutionary games, e.g. the evolutionary trust game, the evolutionary stag hunt, and the evolutionary prisoners dilemma (see, e.g., Binmore, 2007 and Skyrms, 2003).[4] (iii) I am here not distinguishing between *trust* and a *decision to trust* which, for other purposes, might be important to make. Finally, (iv) while the examples offered are meant to be true to our experience of trust, my analysis is not empirical; it is not meant to be a description of real trust complexes, trust as it actually takes place between real agents. Nor is it intended to be a description of the way real trustors attempt to justify their trust if called upon to do so. My focus is rather on the ideal trust complex (of the form of trust at issue) such as might serve as a normative resource for evaluating the degree to which instances of real trust are justified as cases of good judgment.

There are 7 aspects of trust in the trust complex I would like to distinguish, giving each a provisional definition or description of ways that it works as a variable in the trust complex. After introducing these 7 aspects I will discuss some of the issues raised and expand on some of these initial comments.

1. Trust – when S trusts T to A, S makes a *decision* to depend on T; trust seems to create a *dependency* of S on T to A. S's puts herself, so to speak, in T's hands. Such trusting implies that it is not certain that T will A; it is possible that T fails to A and that along with the ability to A, T has the ability to not-A. Thus, in trusting T to A, S *voluntarily* submits to T's power to harm S (more or less severely) by not-A'ing, in the belief – the expectation – that T will A. If this is correct, then it would make sense for S to determine: how *much* should S trust T to A (or: how much should T be trusted to A)? Trust will vary in strength or degrees or amount, and rational trust, then, will always be a rational *degree* of trust rather than trust *simpliciter*. Yet the idea of a "degree of trust" – the idea of a trusting that is less than total – is problematic; it seems that S either trusts T to A, or doesn't trust T to A. It seems odd to say that S partly trusts T to A, say ½ trusts or ¾ trusts T to A. Couldn't T, in such cases, rightfully complain that S in fact doesn't trust T to A – complain that S lacks trust in T? How might the idea of a "degree of trust" make sense? The intuition is that trust is a variable and not all values this variable can take are rational; rational degrees of trust are here proposed to be values that satisfy the constraints imposed by the values other key variables in the trust complex have.

2. Trustworthiness – when S trusts T to A, S believes T to be (to some degree) trustworthy in the sense that S minimally believes that T is willing and able to A. (I include under "willing" such things as: T's being obligated to A, being motivated to A, being taught, trained, programmed or disposed to A, desiring to A, and even pressured/forced to some degree to A. I do not mean to imply that T must have "free will.") The connection between trust and such a belief seems not to be that trust provides the basis for the belief, but rather the other way

around: the belief forms the basis of the trust. Trust is thus *founded* on a belief; take away all belief in the trustworthiness of T to A and S's trust becomes baseless, a kind of unsupported "good-faith" trust. "Trust-no-matter-what" seems much more an act of faith than an instance of trust, or at least an instance of good-judgment trust. If S trusts T to A without *any* belief in T's trustworthiness (something hard to imagine possible) or with unjustified belief (see 3 below) – I will call both "blind trust" – it would have to be examined what reasons might justify such blind trust if and when it occurs. (For example, S might blind-trust T to A as a test of T's trustworthiness, an evidence gathering probe; or perhaps S's blind-trust is a way of teaching T to be (more) trustworthy. In turn, T might intentionally be untrustworthy as a way to teach S to be *less* trusting. Such cases of educational or strategic trust are not, I believe, authentic trust complexes and will not fall under my analysis). Further, if S trusts T to A believing T to be outright untrustworthy to A, then we seem to have a clear case of irrational trust, for such a belief would be the basis for distrust, not trust. Rational trust, then, seems to require a coherence (see 3 below) between the trust placed in the trusted and belief in the trusted's trustworthiness. I will discuss trustworthiness for different kinds of agents below.

3. Trust-supporting belief – if belief in T's trustworthiness is to *support* (as opposed to just motivate) S's trust in T to A, then the belief can't itself be unsupported; it should be justified to S. Thus, S should have evidence *justifying* her belief in T's trustworthiness. It would be too strong, I believe, to claim that S's belief in T's trustworthiness should also be *true* – i.e. that T should actually *be* trustworthy, and that S *knows* this – for it can be rational to believe something false so long as the belief is justified. Also, I believe that default evidence doesn't (and shouldn't) apply; it would be a mistake for S to believe that T is trustworthy to A just because S has no evidence to the contrary, no negative evidence, no reason not to believe T trustworthy and no reason to believe T not trustworthy, as if trustworthiness were the default. Evidence that justifies belief in an agent's trustworthiness will, of necessity, be more-or-less and be subject to updating: that is – inductive.[5] Thus, S's belief in T's trustworthiness will be probable, varying in strength with the amount and quality of the supporting evidence of which S is aware. Here a principle of coherence suggests itself:

T1: A degree of trust is rational (= a rational choice) only if it matches the strength of the evidence the trustor has justifying belief in the trusted's trustworthiness.[6]

It follows from this principle that it is not necessary that the trusted *be* trustworthy for trust to be rational, so long as the trustor is justified in believing the trusted is trustworthy. I am proposing, in other words, that trust is rational relative to S's justified *belief* that T is trustworthy, not relative to T's *actual* trustworthiness. So, for example, where T is very good at deceiving S about T's trustworthiness to A, it could be, by T1, rational for S to trusts T to A. But, by T1, it would *not* be rational for S to blind-trust T to A, where – unbeknownst to S – there exists strong evidence that T is fully trustworthy to A. I note, again, that there might be good reasons for S to trust T to A without sufficient evidence of T's trustworthiness, e.g. to test T's trustworthiness, or to teach T to be more trustworthy, but T1 formally excludes such experimental and strategic "trust" from being cases of rational *trust*, as opposed to cases of experimentation or strategy that happen to involve trust. The possible kinds and sources of evidence that can justify trust-supporting belief will be discussed below for different kinds of T's.

4. Risk – Trust is fragile; there is always the chance that S's trust in T to A is misplaced. If there is no possibility that T does not A, i.e. if T *must* A, then there is no possibility for S to *trust* T to A; for S could be certain that T will A whether S likes it or not. So, given the possibility that T does not A, and might be untrustworthy, in trusting T to A, S risks being let-down, perhaps even betrayed, in the event T fails to A. Trust-risk has been described as trust danger, the trustor being vulnerable to the consequences of the trusted's failing to fulfill the trust place in her (McLeod, 2011). The important point for my exploration is that, as with trust and as with trust-supporting belief in the trusted's trustworthiness, the risk that S's trust will be frustrated or violated *varies*: there can be greater or lesser risk. It would not be correct, I believe, to link trust-risk *only* to T's trustworthiness, as if the less trustworthy T is the more risk S takes-on trusting T to A. It might be that T is very trustworthy to do A, but that circumstances beyond anyone's control, e.g. natural forces, illness, social/political unrest, infrastructure failure, etc., keep T from A'ing. Thus, in determining trust-risk – the probability that T fails to A – S should take into account not only (evidence for) T's trustworthiness/untrustworthiness, but also the chance that the state-of-the-world required for T to A will/will not happen, for T's ability to A is conditional on and relative to the relevant state-of-the-world, it is not absolute.[7] I suggest the following as a link between trustworthiness and risk:

T2: If S trusts T to A, and if state-of-the-world risk is held fixed, then the less trustworthy T is the more risk S has in trusting T to A.

I note that by T2, trust-risk varies relative to the degree of T's *actual* trustworthiness, not relative to S's justified belief in how trustworthy T is. So, if S trusts T to A, it is possible for T to be less trustworthy than S believes T to be (making S too trusting) in which case S risks being let down or betrayed to a greater degree than S might realize. This seems right. But this would not be possible if trust-risk varied *only* with S's (justified) belief in T's trustworthiness, for where the belief is false the degree of trust-risk would still be correct as long as it tracked *believed* trustworthiness as opposed to actual trustworthiness. This seems wrong. Of course, from the point of view of the trustor, in estimating trust-risk S is limited by – *de facto* can only go by – S's (justified) belief in T's trustworthiness.

5. Stakes – if S trusts T to A, then T's A'ing (or the results of T's A'ing) would presumably have *value* for S. It is natural to define the stakes of a trust complex as: the value to S (the trustor) of T's (the trusted's) A'ing (or the outcome of T's A'ing). Minimally, this value is subjective and relative to S, perhaps identifiable with or dependent on no more than S's need or desire that T A's. But I see no reason why what's at stake for S when S trusts T to A can't also be valuable to others who, given that some additional conditions are satisfied, can be considered stakeholders, e.g. groups to which S belongs, networks of which S is part, or even why T's A'ing might not be intrinsically valuable. The important point for this exploration is that T's A'ing (or the result thereof) can be more-or-less valuable to S. Thus, trust-stakes *vary*, and it would make sense for S to determine: how much is at stake in trusting T to A? It is clear, even at this early point, that trust-stakes and trust-risk are linked in how each varies. Even though real agents may fail to connect correctly what they value with the risks involved in achieving or possessing what they value, the classic normative rule in rational choice theory is that subjective value (utility) *should* be reduced proportional to risk; the more risk in achieving an outcome the less that outcome should be valued by the agent desiring it.[8] Because trust is here taken to be a choice on S's part, a principle of rational trust seems to follow:

T3: When S trusts T to A, the degree of trust is rational only if trust-stakes are adjusted downward by the amount of trust-risk.

For simplicity (and following the lead of recent epistemology), the range that trust-stakes can vary will be divided into two convenient segments: *high stakes* (= the value to S of T's A'ing is great, perhaps cannot be substituted), e.g. S's "life-and-limb", or S's basic well-being, or S's livelihood, or things and persons near-and-dear to S, depends on T's A'ing. *Low stakes* (= the value to S of T's A'ing is small, perhaps S can easily trust agents other than T to A), e.g. S loses a small amount of time or money if T fails to A. Principle T3, obviously, becomes more important the higher the trust-stakes.

6. Autonomy – if S trusts T to A, then T must be, to some degree, autonomous – at least in A'ing. Autonomy is, in a sense, already built into the concept of *agent* I am using. If T were not autonomous, if – say – T is under S's complete control, there would be no need for S to *trust* T to A, for S could *make* T do A (thereby avoiding the main problem of trust-risk) and S would then only have to trust himself to make T do A when the time comes. And, if – say – T is under the complete control of an agent (U) other than S, then if S trusts T to A, S is actually trusting U

to make T A. Thus, if S trusts T to A, then S does not directly *force* or *make* T do A (even if S could do so), though by trusting T to A, S might be providing T with a reason, an obligation, or be otherwise motivating T, to A. The importance of T's autonomy in the trust complex is not only positive, it is a double-edged sword. On the one hand, it means T is able to A with "creativity," that T can stick to the "letter" or expand to the "spirit" of S's trust, and that T can fulfill S's trust in a way that best satisfies the variety of limits and constraints within which T might be operating.[9] On the other hand, however, T's autonomy means that T can frustrate or disappoint or betray S's trust, that T has the power to do this, i.e. how willing and able T is to not-A relative to T's willingness and ability to A. T's autonomy, then, is closely linked to T's trustworthiness; namely, the degree of T's *autonomy* becomes a (partial) measure of T's potential untrustworthiness, and conversely T's potential *untrustworthiness* affirms T's autonomy. Autonomy, in this sense, is not the same as the probability that T fails to A because of circumstances beyond anyone's control. It is more the independent ability of T to "decide" not to A, when S trusts T to A. For my purposes, the concept of autonomy from Tavani and Buechner serves well (Tavani and Buechner, 2012b). Three key features of an autonomous agent (slightly re-described in line with my topic) are an agent: (1) capable of acting "on its own" (equivalently – acting causally independently of direct human control); (2) whose actions are meaningful (equivalently – have instrumental value, or contributes to a means-end system, or are appropriate to its environment); (3) where human control typically happens through some form of education, e.g. training, persuasion, learning. Given these three features, it is clear that T's autonomy implies that T can break the trust S places in T to A by not A'ing (and, as an aside, this might be a good thing, e.g. if A = murder, and if T's autonomy allows T to betray S's trust). So, in trusting T to A, T's autonomy is something S should *worry* about, it factors into S's trust risk. Two points immediately follow: first, for T to be autonomous T must have options – minimally two alternatives: to A or to not-A – and T does A (as S expects that T will do) "on its own." Second, as trustworthiness varies in degrees, so too must autonomy. However, while T can be more-or-less autonomous in A'ing, T can't reach the limit of zero autonomy without thereby voiding the possibility of being trusted to A. Provisionally, autonomy would seem to link to risk, to stakes, and to trustworthiness along the following lines:

T4: When S trusts T to A, especially given high stakes, the degree of trust is rational only if the more autonomy T has in A'ing and the more risk S takes-on that T does not A, the stronger the evidence S has justifying S's belief in T's trustworthiness.

7. <u>Options</u> – If S trusts T to A, S's trust could not be a *decision*, it could not be voluntary, if S had no other alternative, no choice but to trust T to A. Trust would be, so to speak, forced upon S. Such might be the case, e.g., if S were in a severe accident, in a state of shock, barely conscious,

and near death; the EMT's are doing their best to stabilize S. Perhaps S in some sense "trusts" the EMT's to do their best to keep S alive, but it would seem that S has no choice in the matter and at any rate, at the moment is not making a decision to place trust in these EMT's. I wish to put aside such cases of "no choice" trust, if there is such a thing, and stay focused on the kind of trust that could qualify (or fail to qualify) as a rational choice. It must be possible, then, for S to ask: what options are available other than trusting T to A? Minimally, there must be the option not to trust T to A, to withhold trust, or to revoke trust if it has already been place in T to A. (It might be worth mentioning, in light of ordinary language use where the statement "I do not trust you" typically means "You are not trustworthy," that S's *not trusting* T to A is not the same as S's *distrusting* T to A, perhaps because S has evidence that T is untrustworthy. And it is not to be confused with S's *mistrusting* T to A, perhaps because S commits an inductive fallacy in forming her trust-supporting belief and ends up trusting T more, or less, than S's evidence justifies.) In the case where S decides not to trust T to A, it does not mean that S believes T cannot be trusted to A. It might be the case that S believes T is very trustworthy to A, but that S has simply decided to do A himself; or perhaps that S has had a change of mind and no longer values T's A'ing (or the outcome of T's A'ing); or perhaps S chooses to trust U to A instead of T for reasons having nothing to do with T's trustworthiness to A.

The main point for my purpose, however, is that given a set of options available to S, one of which is to trust T to A, S makes a rational choice in trusting T to A if this is the best option, but this decision is rational only in one sense or on one level. For once *this* choice is made, S must still confront another set of options, namely, the various possible *degrees of trust* S can place in T to A, from maximal to minimal. Trusting T to A might be the right thing for S to do, but then S could mistake the degree of trust (too much or too little) that should be placed in T to A and end up *mistrusting* T; i.e., S might violate one or more of the above principles T1 – T4. For example, it might be correct that Jill trusts her husband Jack not to cheat when on business trips (let's say that it would be wrong for Jill to withhold this trust, Jack does not deserve complete lack of trust in this regard), and while on business trips Jack in fact doesn't have any extra-marital affair. Yet Jill might place too much trust in Jack to remain faithful, for had the right temptation come along Jack would have cheated on Jill. Or, it might be rational for Jack to trust the computer guided laser robot that will perform his eye surgery, rather than decide not to trust it; but then Jack might get the degree of trust wrong and place too little trust in it, causing him unnecessary worry, nightmares, and discomfort as the surgery approaches.

At this point I will assume that no important variable in the trust complex has been overlooked, that the above 7 captures the key aspects – though clearly not all the significant relationships among them and by no means in the detail and extent that can and should be explored. To review (see diagram below), these are:

(1) the degree of trust the trustor places in the trusted to do something (= A);

(2) the degree of trustworthiness the trustor believes the trusted to have to A;

(3) the strength of the evidence the trustor has justifying belief in the trusted's trustworthiness to A;

(4) the amount of risk the trustor takes-on in trusting the trusted to A;

(5) the value to the trustor of the trusted's A'ing (or the result of A'ing) at stake in trusting the trusted to A;

(6) the degree of autonomy the trusted has to not A;

(7) the 2 levels of options available to the trustor: (i) other than trusting the trusted to A, and (ii) other than trusting to a given degree.

I believe that within this trust complex, given the kind of trust at issue, the idea of a "degree of trust" has received some clarification along with the target idea of a "rational degree of trust" – i.e. that rational trust is a form of *bounded* practical rationality. Additionally, this analysis allows several interesting questions concerning trust to be explored. For example:

a) Where there are multiple trustors (S_1, S_2, S_3, …, S_n) each independently trusting T to A, must each have the same degree of trust in T to A for trust to be rational? No – for at least two reasons: each S might have different amounts of evidence concerning T's trustworthiness, and the stakes might be different for each S.

b) Where trust is sequential, i.e. S trusts T_1 to trust T_2 to A, must S's degree of trust equal that of T_1 for trust to be rational? No – the amount and quality of evidence justifying S's belief in T_1's trustworthiness could differ from that of the evidence justifying T_1's belief in T_2's trustworthiness to A. Also, both the stakes and the risk for S might not equal the stakes and the risk for T_1.

c) Where trust is nested, i.e. S trusts T to A, and agent T, say a legal system, itself contains a trust complex such that agent U trusts agent V to B, and action B is necessary for action A to take place, would S's degree of trust in T to A be rational if U's degree of trust in V to B were not rational? Or, would S's degree of trust in T to A be rational if it were greater than U's degree of (rational) trust in V to B? Without argument, I suggest it would not be in each case.

d) Where trustworthiness is collective, i.e. S trusts T to A, and T = a group of coordinated or cooperating individual agents, perhaps a jury or a committee deliberating some issue, how might (evidence for) the degree of trustworthiness of the collective be aggregated from (evidence for) the different degrees of each individual's trustworthiness?

e) Where trust is distributed over independent trustees, i.e. S trusts T_1, T_2, T_3,…,T_n separately to contribute to A'ing, must S place equal trust in each T? It would appear not, if variables in the trust complex such as trust-risk, trust-stakes, trustee autonomy, and trust-supporting belief are taken into account.

f) Where trust is mediated or transferred, i.e. S trusts T to A because S trusts agent U and U has vouched for (trusts) T, would S's degree of trust in T be rational if it were greater than S's degree of trust in U? Without argument, I suggest it would not.

g) Where trust is reciprocal, i.e. S and T trust each other to A (either simultaneously or in turn) must each trustor's degree of trust be the same to be rational? It seems not, for it is possible that one agent is untrustworthy.[10]

h) Where S trusts T to A, for trust to be rational must A be morally (or legally, or socially) permissible? It seems not, for by the above analysis there can be rational trust among criminals. Then, minimally, must S *believe* that A is morally, (etc.) permissible for S's trust to be rational? Again, within the trust-complex framework as articulated into the above 7 variables it appears not. The moral (legal, etc.) status of A seems to be a worry independent of the question of trust's rationality.

Questions and topics along the lines of the above 8, to the extent that they are worth asking and exploring, suggest that the analysis of the trust complex into the 7 variables presented above might be on-track and fruitful.

Keeping with the trust complex as elaborate along these 7 aspects, I would like now to turn to the question of the kinds and sources of evidence that justifies trust-supporting belief. The problem we face, as I see it, is that when S trusts T to A, the kinds and sources of evidence that justifies S's belief that T is sufficiently trustworthy when T is a *human* agent cannot be used to justify trust-supporting belief when T is a *non-human* agent. If correct, this indicates that trust-supporting belief – and by extension the kind of rational trust founded on such belief – are essentially different for trust-complexes containing human trustees and those containing non-human trustees.[11] In the case of the human trustee, belief in trustworthiness seems to require evidence concerning psychological or moral properties that, in the case of a non-human trustee, don't and can't apply. In the case of a non-human trustee, belief in trustworthiness seems necessarily limited to evidence about T’s *reliability*, which if applied to the case of a human trustee would justify belief in T’s reliability, but not T’s *trustworthiness*. It would appear that for a human trustee the concepts “trustworthy” and “reliable” seem only minimally to overlap, if at all, whereas for a non-human trustee these 2 concepts seem to coincide for low-stake trust and overlap a great deal, if not completely, for high-stake trust.

An alternative, perhaps equivalent, way to put the distinction is this: when T is human and S trusts T to A, it is possible for T to *betray* S’s trust, and it is possible for S to *distrust* T to A. And, importantly, it is possible for T to betray S’s (high stakes) trust so profoundly that it does not only reveal T’s untrustworthiness, it has the power to “implicate” S and affect S’s own moral phenomenology: perhaps it undermines S’s moral self-confidence, or weakens S’s sense of moral integrity, or motivates in S deep self-blame, or otherwise diminishes S in her own eyes. Rebuilding trust, once trust has been betrayed, becomes next to impossible as trust-stakes go up, not only because the damage to the trusted’s reputation becomes at some point irreversible but also because the damage to the trustor’s self might be irreparable, leaving S incapable of trust in complexes where T should be trusted. In contrast, when T is a non-human agent and S trusts T to A, none of this seems possible: T can’t betray S’s trust, nor can S distrust T. But, importantly, when T is a non-human agent it does seem possible for S to *mistrust* T to A, and when this happens rebuilding trust, even in high-stake complexes, is not only possible, it seems a relatively easy thing to accomplish (in principle if not always in practice); neither S nor T are irreparably damaged.

First, take the case where T is a human in whom trust has been placed to A. What kinds of evidence would normally (typically) be available to S that would justify S’s belief in T’s trustworthiness? How might T “earn” or “deserve” S’s trust? I see 4 possible sources:

1. Evidence from personal relations – perhaps T is linked to S by feelings of affection, bonds or pledges of loyalty, family ties, or T emotionally cares about S strongly enough to assure S that T can be trusted to A. That T will A out of emotional attachment, loyalty, care, family connections, etc. provides pretty strong evidence for S’s trust-supporting belief that T is trustworthy.

2. Evidence from self-interest – perhaps there are social norms, cultural standards of behavior and promise-keeping, legal or religious obligations, contractual roles, or an expectation of professionalism that T lives by and that motivates T to be trustworthy, such that it would reflect badly on T in the eyes of others, perhaps even land T in trouble, or cause others to break T’s trust in them, should T fail to fulfill S’s trust; or such that it would bring rewards of honor or

advancement should T prove trustworthy. T's self-interest in this case would be evidence for S to believe in T's trustworthiness to A.

3. Evidence from character (virtue) – perhaps T possesses the general virtue of trustworthiness (typically along with such virtues as honesty, being responsible, having perseverance and commitment, etc.); it is one of T's strengths of character, a quality of T's personality or make-up, as e.g. revealed in the history of T's interactions with others as a trustee, built into T's reputation, and described in T's letters of recommendation. T's virtuous character would serve as good evidence for S's trust-supporting belief that T can and should be trusted to A.

4. Evidence from morality – perhaps T is motivated by moral reasons, by principles of moral duty; trust having been placed in her, and accepting S's trust to A, T understands that being trustworthy is the right and good way to be. If S knows that T is a morally principled person, S has pretty strong evidence, absent any clash of moral duties that T might be experiencing and absent T's believing A to be morally wrong, to believe that T can and should be trusted to A.

These sources are easily recognize as the personalist, consequentialist, Aristotelian, and Kantian "takes" on the kinds of evidence that might justify belief in T's trustworthiness relative to A'ing.[12] It is worth noting that, where T is a human agent: (i) these 4 sources could also provide evidence for T's untrustworthiness; (ii) for each source, evidence can be more-or-less, justifying a stronger or a weaker belief in T's trustworthiness; (iii) any 2 or more of these sources can combine or conflict to form stronger or weaker evidence for a trust-supporting belief in T's trustworthiness (e.g. the classic business manager's dilemma: employee T might be very professional in fulfilling work obligations (= trustworthy), but might also have strong feelings of ill-will toward manager S (= untrustworthy); what should S do, trust T to A or not?); (iv) some of these sources, e.g. 1, requires that S not be anonymous to T, while others, e.g. 3, allow S and T to be anonymous to each other; (v) none of these sources require that S and T be contemporaries.

Now take the case where T is a non-human agent, a smart "machine" in Tavani's sense of the term (Tavani, 2012a, Ch. 12). For example, S trusts T to A where S = airline passengers, T = the plane's automatic pilot, and A = land the plane safely is dangerous weather conditions; or, S = eye patient, T = computer guided/controlled laser eye surgery system, and A = remove cataracts without further damaging vision; or S = the owner-passenger, T = a fully autonomous smart car, and A = drive S and her family through Los Angeles traffic to an important doctor's appointment. These are high stake trust complexes. (I leave aside the interesting case of a blind person trusting a seeing-eye dog; the trusted is a non-human agent, but not a smart "machine" and might well fall under the 1st or 2nd category of evidence for human trustworthiness. Trusting an animal seems to be midway between trusting a human and trusting a smart-machine.) Keeping to high stake trust I will use this smart-machine scenario: T = IBM's Watson, S needs a very expensive medical procedure without which there is a high probability, though not a certainty, that S will have serious chronic ill-health and perhaps even die. Any monetary winnings Watson gains playing *Jeopardy* will go towards S's medical costs, potentially paying them in full. If Watson loses at *Jeopardy*, penniless S can't receive the medical care that is needed. In this scenario, S trusts T to A = S trusts Watson to win at *Jeopardy*, and what is at stake is S's health and probable continued life. What evidence could S have that justifies belief that T can and should be trusted to A? It is obvious that none of the above 4 sources where T is a human agent can be applied to Watson. Watson is not (1) personally or emotionally attached to S, or (2) fulfilling its role of trustee out of self-interest, or (3) acting out of a virtuous character, or (4) doing its duty because it is the morally principled thing to do. The problem is not that S believes Watson is untrustworthy, and therefore should be *distrusted*, because S has evidence that (1) Watson is ill-disposed toward S, or (2) it is in Watson's self-interest to fail to A, or (3) Watson's character is full of vices, or (4) Watson is morally unprincipled. The problem is that it seems impossible for there to be any such positive or negative evidence about T's trustworthiness where T is a smart-machine – the kind of evidence that is typically demanded when T is human. (It is perhaps interesting to note that in the science-fiction genre, at least as far as I'm aware of it, smart-machines are typically given "personalities" falling under source 3: they have a virtuous or a vicious character as a matter of "wiring/programming" in somewhat the same way that some thinkers have theorized that humans are "naturally" or "innately" of good or bad character.)

What kind of evidence, then, is available to S to justify a trust-supporting belief that Watson can and should be trusted to win at *Jeopardy*? The only source that I can think of is Watson's past performance playing *Jeopardy*, the record of experiments in which Watson was tested competing at *Jeopardy* against strong human and smart-machine opponents. Had Watson never been tested at *Jeopardy*, all the testimony in the world of expert smart-machine designers and builders would not be sufficient to justify belief that Watson can and should be trusted to win playing *Jeopardy* against strong human players – these would be no more than unverified predictions meeting with deep skepticism (see mediated trust (f) above). It would be "blind trust" if in the absence of a record of past performance S nevertheless trusts Watson to win at *Jeopardy*; it would not be rational trust. The same applies to the above examples of the auto-pilot landing the plane, the laser eye-surgery robot removing cataracts, and the smart-car driving a family through traffic; evidence justifying belief in the trustworthiness of these non-human agents (smart-machines) would have to come from the record of past performance under rigorous testing conditions, i.e. that they are "tried and true." I can't imagine belief in a machine's trustworthiness being justified by a different kind of evidence (however this might be my own limited imagination).

Now what kind of evidence is this? The expectation of future performance based only on a record of past performance seems "thin" compared to the 4 sources in the human agent case. If T were a human agent, such past-performance evidence could on its own justify belief only in T's reliability, not T's *trustworthiness*. With enough evidence from successful past performance, S is justified to believe that T is *dependable*, that S can rely on - depend on - T to A. For a human T, however, something more –something *different* – seems required to (justifiably) attribute trustworthiness than is required to (justifiably) attribute dependability. In fact, I can see no reason why, where T is human, it can't be rational for S to trust T to A when T has never A'ed before (none of the principles T1 – T4 require past performance A'ing); but where T is a smart-machine such trust would be baseless. Why is this? The human case seems to require evidence that something central to the *person* of the trusted – something normative, and perhaps even metaphysical in the sense of (partly) constitutive of T's *self* – carries over from the past or the present into the future; it does not require evidence that successful past performance carries over (this evidence would be, so to speak, icing on the cake). The human case seems to require that there is a core of psychological or moral *stability* to the person, i.e., a re-identifiable *self*, or quality thereof. This is what the 4 sources of evidence for human trustworthiness seem to have in common. The smart-machine case, however – the case where evidence *must* address past success – requires only high statistical probability that past and present dependability will continue in the future. This is a functional-norm requirement, not a person- or moral-norm requirement, and is the *only* kind of evidence that seem available and applicable to justify trust-supporting belief in the case of a smart-machine's trustworthiness.

There is a related distinction between a human and a smart-machine trustee whose significance, I believe, is less than it might at first appear. Where T is human and S trusts T to A, it seems impossible for T not to know (or believe) that she is being trusted to A (although S might be anonymous to T); it would be something else, not trust, if T had no idea whatsoever that she is being trusted to A when she A's. If this is correct, then *both* S *and* T believe that T is to some degree trustworthy; that is, in accepting the trust placed in her, T *also* believes that she is sufficiently trustworthy to fulfill the trust placed in her to A (excluding the case in which T is deceiving S). In other words, if S trusts T to A, and T is human, then T must have it would seem (some) *self-confidence*, (some) *self-trust* to A. What evidence does T have that justifies his belief that he himself is sufficiently trustworthy to A? It would presumably be the same evidence that S has to justify S's trust-supporting belief in T's trustworthiness (plus perhaps additional evidence unavailable to S, e.g. S might believe that T is virtuously trustworthy, but might not know that T also cares for S, whereas T presumably knows relative to A'ing both that she has a trustworthy character and that she cares for S). It would be an odd and unusual trust complex if S trusts T to A *more* than T trusts herself to A. Nevertheless, in the case where the trusted is human, the trusted has at least *some* relevant self-knowledge (self-awareness, self-confidence), and this self-knowledge it would seem is partly constitutive of the trusted's trustworthiness. A smart-machine, however, has no such self-beliefs or self-confidence, and so – it might be argued – the trustworthiness of such a machine cannot be significantly different from its reliability, whereas human trustworthiness is importantly different from (mere) human reliability. This explains why, for example, a smart-machine can be *made* (designed) to be trustworthy (i.e. reliable), but a human can't be similarly made to be trustworthy but must learn or be taught, or be otherwise motivated to be.

The problem with this line of reasoning, I believe, is not so much the fact that human self-knowledge is deeply mysterious and philosophically controversial, it is that smart-machines have self-referring abilities of great power and sophistication. They clearly have well-tested self-monitoring, self-controlling, self-correcting and learning abilities – the smart-machine equivalent of human self-confidence – without which they couldn't be trusted, especially in high-stake trust complexes. What is wanted, pragmatically speaking, from a human trustee's self-knowledge are those things that make and keep the person trustworthy. I can't see much difference here, pragmatically, between a human trustee's self-beliefs and self-control, and a smart-machine's "self-awareness" and self-control with respect to trustworthiness.

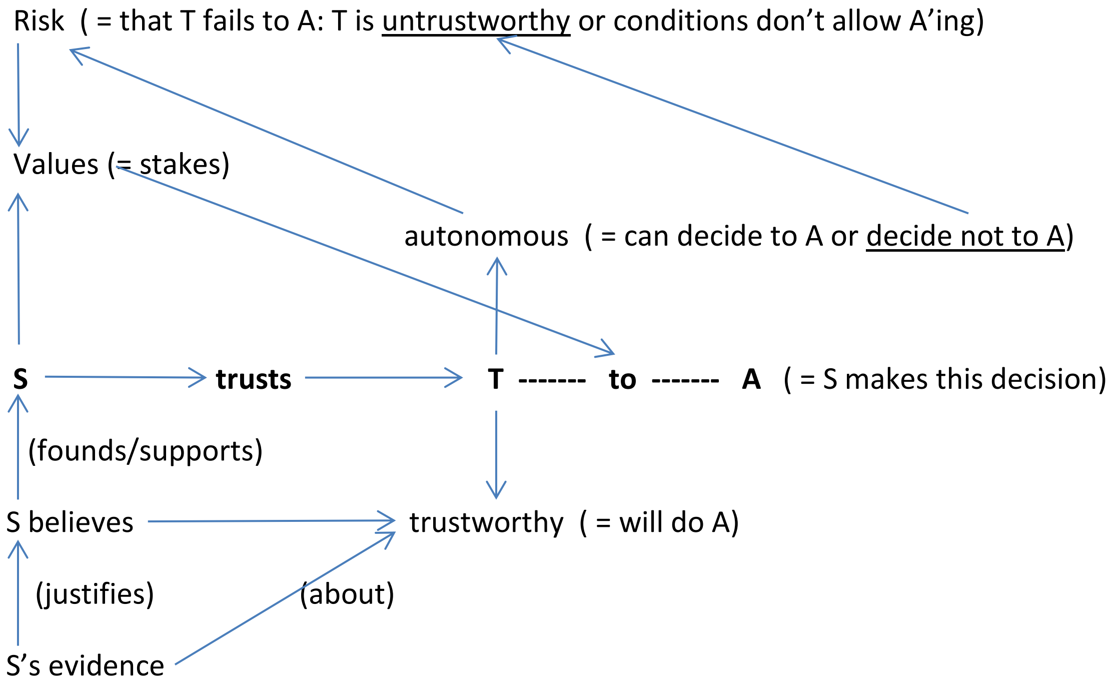
If the distinction that I have argued for between human trustworthiness and smart-machine trustworthiness within the trust complex is correct, I conclude with these proposals:

1. In low-stake trust complexes, if the trusted is a smart-machine, there seems to be no difference between trustworthiness and dependability (reliability), and no difference between the trustor's trust-supporting belief in the trusted's trustworthiness vs. its dependability. The kind and degree of trust in each case is the same and the amount and quality of evidence justifying trust-supporting belief can be minimal.

2. In high-stake trust complexes, if the trusted is a smart-machine (or network of such), the difference between trustworthiness and dependability seems to be one of degree not kind; trustworthiness = a high degree of trustee dependability coupled with a high degree of trustee autonomy. In such a complex trust seems to be linked more to the trusted's autonomy and reliance more to its dependability, but there appears to be no essential difference between trust and reliance. The amount and quality of evidence justifying trust-supporting belief in trustworthiness, by principle T4, should be stronger than that justifying trust-supporting belief in lower-stake dependability, the higher the stakes and the more autonomy the better the evidence for trustworthiness from past performance should be.

3. In both low- and high-stake trust complexes, if the trusted is human, the difference between trustworthiness and dependability seems to be one of kind not just degree. To justify belief that the trusted is sufficiently *trustworthy* – that is, for trust to be rational, and that the trusted can and should be trusted – seems to require evidence concerning certain enduring qualities of the person(s) trusted, evidence concerning the kind of *fixed* person s/he (or they in the case of groups, teams or networks of persons) is from a psychological or moral point-of-view, even in cases where the action S trusts T to do is itself morally, legally, or socially, etc. impermissible. In contrast, evidence concerning human dependability seems to have no different requirement for the trusted human as for the trusted smart-machine.[13]

Diagram of the trust complex: *S trusts T to A*. Arrows represent relationships among the main variables as these have been described and argued for in the text above.



Risk (= that T fails to A: T is untrustworthy or conditions don't allow A'ing)
Values (= stakes)
autonomous (= can decide to A or decide not to A)
S
trusts
T ------- to ------- A (= S makes this decision)
(founds/supports)
S believes
trustworthy (= will do A)
(justifies)
(about)
S's evidence

Notes:

[1] The kind of analysis I am attempting is eidetic analysis, associated with phenomenology, and specifically with Husserl's analysis of the structures of consciousness. Briefly, eidetic analysis describes a typically idealized structure or complex – in this case: trust – so as to isolate its significant parts (its "moments" in the terminology of phenomenology), and subjects these to imaginative variation in order to discover necessary ("essential") relationships among them of, e.g., compossibility, dependence, independence, supervenience, and degrees of co-variation. If (for the moment) we assume that the object of a concept is an essence rather than an extension, eidetic analysis can be thought of as a kind of multi-variable conceptual analysis rather than akin to the methods of linguistic or ordinary language analysis.

[2] Tavani and Buechner offer 5 conditions, but in their analysis focus primarily on the normative element in what they refer to as the "trust relation."

[3] This way of representing the kind of trust with which I will be concerned with has become fairly standard in the literature. See, e.g., MacLeod 2011, and Ullmann-Margalit 2002.

[4] The literature on trust and related complexes of cooperation, reciprocity and mutual altruism – or rather the evolutionary dynamics that displaces their role – within the framework of evolutionary (as opposed to rational choice) game theory ranges from the highly technical to the popular. Axelrod (1984) on the prisoner's dilemma has become a classic. See also Binmore (2005) especially Ch. 5 on trust and reciprocity, and Skyrms (2003) on the problem of trust and cooperation in the stag hunt.

[5] We might consider, here, the possibility of non-inferential justification. As I see it, there are two forms this could take: (i) perhaps S's belief in T's trustworthiness is self-evident and thus self-justifying, akin to certain basic principles such as "whatever is, is." Or (ii) perhaps S has a direct non-propositional insight or intuition of T's trustworthiness, a "view into T's soul" as it were. The problem with non-inferential justification, at least in these two versions, is that they are too strong; they almost guarantee T's trustworthiness, in the same way that S's *knowing* that T is trustworthy would. Trust, however, can't be founded on such a strong base for it is not robust; given the ease and frequency with which trust can be (and is) betrayed, its fragility seems to require a foundation in something weaker: a belief whose inferential justification makes it probable, not certain.

[6] See section 4 of MacLeod for a comparison of some epistemic conditions suggested for trust, and for possible criticisms from the view point of non-epistemic trust.

[7] deVries (2011) suggests a form of trust that is unconditional. But unconditional trust, if such a thing is possible, would presumably not have the form: S trusts T to A.

[8] Of course, this does not mean that the objective value of something an agent desires to achieve should decrease as risk of failure to achieve it increases. So, for example, if person A desires to marry person B and has a high probability (risk) of being rejected, the rational choice norm in question isn't suggesting the objective (intrinsic) value of B as a person or of the institution of marriage should be discounted (weighted) by the rejection probability.

9. I leave aside the difficult but important problem of identifying which actions count as A'ing, for any action type A. In the theory of action, an action is closely connected with an intention and with a description. If S trusts T to A, is S's trust fulfilled or violated if T intends to not A and to do B instead, but as it happens action B can also be described as an instance of A'ing? For example, suppose S lends T money and trusts T to A = repay the loan by a certain date; T meanwhile has no intention of repaying it (intends to not-A) but plans to give S counterfeit money instead (= B). When the time for repayment comes, T mistakenly gives S the wrong envelope, the one containing real money, and keeps the wrong envelope, the one containing the counterfeit notes. T has "given S money," but has T fulfilled S's trust to A (repaid the loan) or not? Similarly, suppose S trusts T to attend and to applaud S's piano recital performance (this means a great deal to S), and T attends but decides S has made so many elementary mistakes in her playing that she won't applaud; however, while others are applauding S's performance T energetically claps her hands together several times in an effort to kill an annoying insect flying near T's head. Has T fulfilled or disappointed S's trust? The same motion "clapping hands together" could count as "applauding" or as "killing an annoying insect" depending on several conditions, including the intention of the agent and the description under which the motion is done. The problem of identifying what counts as A'ing, when S trusts T to A, becomes even more puzzling when T is a smart machine and A'ing is based on the mechanical motion of running an algorithm or program.

10. Reciprocal trust has been widely explored within the framework of rational choice theory, but less as an individual decision than as a strategic decision within the structure of a game; see, for example, Ulmann-Margalit (2002) for an analysis of reciprocal trust as a rational game; by way of comparison see Skyrms (2003) for a reduction of reciprocal trust to dynamic movement toward cooperative equilibria within evolutionary games.

11. I base this claim on two principles that seem intuitively correct but are not argued for here, though their positive versions are based on the connections among the 7 variables in the trust complex as analyzed above. Suppose we have 2 trust complexes such that evidence set (E_1) justifies trust-supporting belief (B_1) which founds an instance of trust (t_1) in one, and evidence set (E_2) justifies trust-supporting belief (B_2) which founds an instance of trust (t_2) in the other (and this is all we have), then (1) if E_1 *could not* justify B_2 and if E_2 *could not* justify B_1, then $B_1 \neq B_2$ and B_2 *could not* found t_1 and B_1 *could not* found t_2; and (2) if B_1 *could not* found t_2 or if B_2 *could not* found t_1, then $t_1 \neq t_2$ and the kind of trust of which each is an instance are different in essence. What will be argued is the antecedent of (1); given these 2 principles, the consequent of (2) should follow.

12. I am assuming here that *evidence* justifying belief in T's trustworthiness is to be distinguished from the *assessment* of T's trustworthiness. See, for example, chapter 8 in Kahneman (2011) for studies of facial features and related visual "heuristics" that appear to motivate rapid, automatic, emotionally linked assessments of trustworthiness.

13. This study has benefited from helpful and challenging comments from my colleagues Jerry Dolan, Herman Tavani, and John Caiazza at our Nashua Circle reading. Special thanks to Herman Tavani for many stimulating discussions about the philosophy of trust.

Bibliography:

Axelrod, Robert. 1984. *The Evolution of Cooperation*. Basic Books.

Binmore, Ken. 2005. *Natural Justice*. New York: Oxford University Press.

Buechner, Jeff, and Herman T. Tavani. 2011. "Trust and Multi-Agent Systems: Applying the 'Diffuse, Default Model' of Trust to Experiments Involving Artificial Agents." *Ethics and Information Technology* 13, no. 1: 39–51.

deVries, Willem. 2011. "Some Forms of Trust." *Information* 2, no. 1: 1-16.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

McLeod, Carolyn. 2011. "Trust." In E. Zalta, ed. *The Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/archives/spr2011/entries/trust/.

Skyrms, Brian. 2003. *The Stag Hunt and the Evolution of the Social Structure*. Cambridge: Cambridge University Press.

Tavani, Herman T. 2012a. *Ethics and Technology; Controversies, Questions, and Strategies for Ethical Computing*. 4th ed. Hoboken, NJ: John Wiley and Sons.

Tavani, Herman T, and Jeff Buechner. 2012b. "Autonomy and Trust in the Context of Artificial Agents." In M. Decker and M. Gutmann, eds. *Evolutionary Robotics, Organic Computing, and Adaptive Ambience*. Berlin, Germany: Verlag LIT.

Ullmann-Margalit, Edna. 2002. "Trust out of distrust." *The Journal of Philosophy*, Vol. 99, No. 10: 532-548.