Trust – distinguishing forms, kinds, and degrees: I

Lloyd J. Carr
Philosophy Department
Rivier University
Nashua, NH 03060
lcarr@rivier.edu
Website: http://www.rivier.edu/faculty/lcarr

Part I: Kinds of trust

Abstract

This paper is part I of a three-part project to distinguish three central features of trust: that trust has different *kinds*, *forms*, and *degrees*. In the first section of this paper, I describe (on an intuitive level) examples of degrees, kinds, and forms of trust, and by stipulation assign the terms "degree," "kind," and "form" for these different ways that our trust can and does vary. I argue that in the trust literature these three features of trust (under a wide variety of names) are not sufficiently differentiated, and that this constitutes a gap in our understanding of trust. In the second section, I set up a general trust framework, a basic unit of trust, within which degrees, kinds, and forms of trust might be distinguished and explored. This general framework of trust is presented and argued for in somewhat more detail than necessary for the purpose of this paper, because it is intended to be re-used in parts II and III of my larger project. In the third section of this paper I focus on *kinds* of trust. As an alternative to the standard strategy in the trust literature of basing kinds of trust on kinds of agents as potential receivers of trust, I propose we start with kinds of trust – ways that a human agent forms a relation of trust with different kinds of potential trustees – and describe three: communication-based trust, practice-based trust, and context-based trust. I examine the structure of each, and argue for a set of principles/conditions for each. On the basis of these kinds of trust, I argue for certain categories of agents as possible trustees for each kind. I end this paper with a comparison of the concept of context-based trust with a similar "contextual theory of trust" developed by Buechner and Tavani (2011), finding mutual support as well as some differences.

Keywords: Agents, Human agents, Artificial agents, Practical agents, Trust, Trust kinds, Communication, Practice, Cooperation, Context, Network structure

1. Introduction: a problem

Anyone who has trusted or has received trust from another person has experienced certain basic possibilities about trust, possibilities that appear to be so much a part of trusting that no giver or receiver of trust would, upon reflection, doubt their importance or deny their existence. There are three such possibilities I wish to describe and use as "anchors" or "fixed points" in the sense that they should not be left unexplained by our concept of trust, and also in the sense that they can serve as guides or "checks" in developing a philosophical understanding of trust.

The first is that *it is possible to be too-trusting or not trusting enough*. This means that in trusting another, or in being trusted, there is an "amount," "degree," or "strength" of trust that is experienced (or perhaps in retrospect evaluated) to be more-or-less "right," "appropriate," or in some sense "justified," and an "amount" or "degree" or "strength" of trust that is experienced (or perhaps later judged) to be "wrong, " perhaps "too much" or "too little" or in some sense "unjustified." That trust can vary along the dimension of "intensity" or "amount" or "degree" or "strength" is a basic possibility for which any concept of trust must account. Let variations within this possibility be called: **degrees of trust**. With respect to degrees of trust, I take the central question to be: When we give or receive trust, what determines the "right" degree of trust?

The second anchor point that our trust-experience reveals is that *it is possible to trust a variety of agents, but we don't trust them all in the same* way. These agents fall into different categories and sub-categories. We typically trust *individual human agents*. They might be, for example, experts in a particular field of activity such as medicine or law; the one trusted might be a family member, a friend, or a neighbor; or they could be anonymous officials, e.g. someone who safety-inspects the food we eat. We also trust *institutions* such as a business that manufactures a product or provides a service we use, or a government agency that regulates the distribution of an important resource. We trust *processes* such as a legal trial we might be going through, or the annual performance-review of the job we are doing. Increasingly, we find ourselves trusting our technology, for example, its *smart-machines* and *intelligent non-human systems* such as the automatic pilot that lands our plane in foggy weather conditions, or the "automated-teller" that takes our bank deposits or gives us cash and then correctly up-dates our financial records, or the automated micro-organism sensing system that safety-inspects the food we eat, or the surgical robot that will perform the operation you or a loved-one needs. If

reports are accurate, in the near future we will trust our completely self-driving car to deliver us safely at our destinations. *Domestic animals* form yet another important category of agents to which trust is given: a blind person trusting his seeing-eye dog, a lost and disoriented equestrian trekker trusting her horse to return safely to the stable. That trust can vary according to "the kinds" of agents we trust is a basic possibility for which any concept of trust must account. Let variations within this possibility be called: **kinds of trust**. With respect to kinds of trust, I take the central question to be: In what ways should trust be responsive to the different kinds of agents to which we give our trust?

The third anchor point that our trust-experience reveals is that *it is possible, in trusting another, to be trusted in return*. For example, in trusting your peers to be honest in evaluating your reasoning as an academic, your peers might likewise trust you in the same respect. But it is possible that you are not trusted back in the same respect, but in some other respect; for example, you trust your spouse to point out your deficiencies as a parent, and your spouse trusts you to pick up your child from kindergarten. And it is also possible that, in trusting, you are not trusted in return at all, as would be the case when you trust the automatic pilot to land your plane safely; such an expert system would not be capable of trust. That trust can vary in possible patterns of "reciprocity" or "non-reciprocity" is a basic feature for which any concept of trust must account. Let variations within this possibility be called: **forms of trust**. With respect to the different forms of trust, I take the central question to be: How are the different forms of trust related to different kinds and different degrees of trust?

Given that these three possibilities – that trust has (keeping to the terms I've introduced) *forms*, *kinds*, and *degrees* – are central to the nature of trust, it is important for our understanding of trust to distinguish them carefully and to clarify the ways they might be related; confusion or neglect with respect to these three features would constitute, it would seem, a serious gap in our understanding of trust. In the mainstream literature on trust each of these three features receives recognition and various levels of analysis, as one would expect of important features of any object of study. Yet, (to my knowledge) in this literature we find (i) no commonly accepted terminology by which to refer to these features, (ii) no systematic presentation of their differences, and (iii) no analyses or theories that attempt to explain their interconnections. Concerning these three features of trust, Russell Hardin's (2002, xxi) general claim is apt: "Unfortunately, conceptual issues in the understanding of trust are far messier and more complicated than one might hope. Clearing up these issues turns out to be a major task." The following illustrates the current "messy" state in the mainstream literature.

Carolyn McLeod (2011, sec. I), in her extensive survey of the trust literature, describes analyses and theories in which trust varies in: (a) "amounts" (more or less) according to the "level" of risk/vulnerability to which the trustor is exposed; (b) "types" or "kinds" according to the

presence or absence of optimism on the trustors part; (c) "forms" that are "usual" or "unusual"; and (again) (d) "amounts" according to the confidence the trustor has in the trusted's competence, motivations, or commitment. Annette Baier (1986), in her influential *Ethics* article, refers to "varieties" of trust and lists moral and immoral trust relationships, and states the need to distinguish "forms" of trust according to their "morally relevant features." She mentions "ways of trusting" as linked to "whom one trusts," and in reviewing inquiries into "forms" of trust by past philosophers refers to Aquinas' writings on trust in God and Locke's on trust in government and officials. In her Tanner Lectures (1991, Lecture I), Baier seems to forego the classificatory language of "varieties," "ways," and "forms," but does introduce the idea of "sorts of trust" that are to be distinguished along the lines of trust between intimates, personal trust, and impersonal trust. Turning to the work of another important contributor to the literature on trust, Hardin (2006, 18) states, *"…once we have relevant knowledge – of your moral commitments, of your psychological or character disposition, or your encapsulation of our interests – that knowledge constitutes our degree of trust or distrust.*" linking "degrees" of trust with the trustor's knowledge of the trusted. Elsewhere (2002, 21ff) Hardin distinguishes "levels" of trust according to the "level" of the trusted: trusting an individual, trusting an institution, trusting government, trusting society. Diego Gambetta (1988, 213-237), in his summative essay to his important anthology on trust, refers to "intensities," "degrees," and "levels" of trust that depend on and vary with such diverse matters as "the subjective probability with which an agent assesses that another agent or group will perform a particular action" (p. 217), degree of freedom of the trusted (p. 218), social mechanisms that enforce cooperation, constrains on the trusted, risk to the trustor, and interests of the agents (pp. 220-222). Trudy Govier is another major contributor to our understanding of trust. In her "Dilemmas of Trust" (1998, p. 131ff) she discusses the problem of "more-and-less" trust: how much trust is "enough," "too much," or "too little," and argues that such variations depend on what makes trust "reasonable and right." Govier goes on to argue that both trust and distrust, are "susceptible to degrees," and names three: "slight," "moderate," or "complete."

In this sample of some of the most influential literature on trust we see a remarkably wide range of terminological freedom and conceptual flexibility. On the one hand this seems normal and appropriate given that: (i) these thinkers approach the topic of trust from different philosophical and research methods, and (ii) trust is a human relationship whose importance, complexity and variety is becoming increasingly recognized by researchers. Such freedom and flexibility are valuable, perhaps necessary, for making initial conceptual inroads. But, as the sample above shows, it is also the case that a body of conceptual insights in the literature of a given subject (e.g. trust) can be a potential source of confusion if left, as a field, unsystematic and without a unifying level of terminological and conceptual consistency.

If we look outside the mainstream trust literature, we see a similar diversity; within each work one finds the expected precision in the use of key terms and concepts, but among these works there is little terminological and conceptual uniformity. Willem de Vries (2010), for example, distinguishes three "forms" or "species" of trust on the bases of an analysis of the surface grammar of English expressions: "*trust that …*," "*trust … on …*," "*trust in …*," "*trust … to …*," "*trust … with …*," and simply "*trust….*" Stephen March (1994) is centrally concerned with reducing terminological and conceptual ambiguity in the study of trust, largely because his project is to "formalize trust" to a level sufficient for computational use, a "formalism for trust" that can be "implemented" by an "artificial reasoning agent" (i.e., trust as a variable in such an agent's decision making). March distinguishes "aspects" of trust: conceptual, social, and biological; again "aspects" of trust: basic trust, general trust, situational trust; and a variety of "values" of trust: (a) values representing increases/decreases in trust as a function of amounts of cooperation, costs, and benefits that takes place between agents, (b) values representing degrees of the presence of trust (trust, blind trust, no trust, distrust), and (c) values representing how trust varies with situational utility. For a third example, D. Harrison McKnight and Norman L. Chervany (1996), in the *Abstract* to their wide-ranging meta-study of theories of trust, state, "*Conceptual confusion on trust makes comparing one trust study to another problematic. To facilitate cumulative trust research, the authors propose two kinds of trust typologies: (a) a classification system for types of trust, and (b) definitions of six related trust types that form a model.*" Their "types of trust," however, are actually not types of trust but rather types of theories of trust, a "classification system" of concepts of trust as opposed to trust itself.

Such terminological and conceptual diversity, in Hardin's words, "… messier… than one might hope … ," justifies an effort to distinguish forms, degrees, and kinds of trust, on the basis of which a systematic exploration of how these three features co-vary and interconnect becomes possible. In this paper, I focus only on *kinds* of trust, provisionally defined as: different ways we trust different agents. This intuitive meaning gains plausibility and precision when placed within a general trust framework, the basic unit of trust presented below in section 2, and when three kinds of trust are distinguished and analyzed in section 3. The content of what follows is abstract in so far as kinds of trust are examined in isolations from the closely related features of degrees and forms that would co-occur in real instances of trust. It is also concrete in so far as it is primarily a descriptive exercise rather than an exercise in argument. In two follow-up papers, I plan to take up the remaining two trust features: forms and degrees.

2. Trust: the basic unit

The base unit I will be using is a standard model in the literature on trust:[1] trust takes place within a *3-term relational complex* – a complex state-of-affairs within which three essential parts have (or take on) certain functional roles and in doing so establish certain relationships to each other. The parts and their functional roles are: (i) an agent (symbolized by "S") that functions as the *trustor*, the source of trust; (ii) another agent (symbolized by "T") that functions as the receiver of S's trust, the *trusted* or *trustee*; (iii) an *action*, a series of actions, or a type of behavior (symbolized as "ϕ") that functions as the thing T is trusted to do. In symbols: (*S trusts T to* ϕ). Because S, T, and ϕ are essential to the unit of trust that I am examining, a trust-complex can't exist without these three parts, each functioning in the described way. There are three pairwise relations that hold in the basic trust complex when these three parts function in these three ways: (1) a trust relation between S and T, (2) a value relation between S and ϕ, and (3) an executing relation between T and ϕ. A general picture of trust will emerge as I examine each of these relations.

2.1 The trust relation

The relationship that holds between S and T is ***trust***. As a binary relation, trust has three standard relational properties:
(a) it is non-reflexive (in trusting T, S might or might not have self-trust),
(b) it is non-symmetrical (in trusting T, S might or might not be trusted by T), and
(c) it is non-transitive (if S trusts T and T trusts agent U, S might or might not trust U).

Trust, we see, is similar in these three relational properties to the familiar binary relations "likes" and "loves," but not similar to the binary relations "identical to" or "as tall as." We also see that the trust relation between S and T is directional; it runs from a trustor (S) to a trustee (T), not the other way (even though T as a potential trustee might have elicited or motivated S's trust).

For the purpose of this paper, I define the trust relation to be:
(a) a *decision* an agent (S) makes about another agent (T),
(b) to *expect* this other agent (T) to do an action (ϕ),
(c) based on S's *belief* that T is sufficiently trustworthy with respect to ϕ'ing.

Some brief explanatory comments are in order. Trust, under this conception, is primarily cognitive; it involves a decision, an expectation, and a belief that a human agent (S) forms relative to another (human or non-human) agent (T).[2] Trust is not being defined as a behavioral disposition, as a character trait (as we would think of a virtue or a vice), as a mental attitude

associated with a type of personality (such as a person described as an "optimist" or as a "serious type"), or as an emotional state. I don't mean to deny that a trusting agent won't exhibit trust-related patterns of behavior in triggering situations, have a certain character profile, be a type of personality, or experience emotional states; it is just that trust is not here being defined in terms of these.[3]

We see that the above definition makes trust an intentional, a deliberate, relationship a human agent (S: the trustor) establishes with another agent (T: the trustee); trust is not the kind of relationship that happens involuntarily or automatically as in the case, for example, of being the son or daughter of your parents, or being a citizen of the country in which you happen to be (legally) born. Condition (a) – that trust requires a decision – is meant to capture this idea that the trust relation is established by a consciously performed set of mental acts of a *practical agent*: the trustor. By the above definition, the trust relation cannot be established by T, the "trust-patient," nor must trust be the result of a cooperative agreement between the trustor and the trustee.

I also note that this definition assigns to each instance of trust *meaning* or *mental content* on the part of the trustor; as a set of mental acts (decision, expectation, and belief) this content is primarily conceptual and in part propositional; but trust itself is a relation between agents, it is not being defined here as a propositional attitude. Because trust has conceptual content, trust is a *meaningful* (i.e. semantic) relation (not a physical relation): one that is subject to norms of rationality and to critical reflection on the part of the trustor and others, and one that can have meaningful (semantic) links to the contents of other cognitive activities in the trustor's life; these would be difficult to understand, if not impossible, under some non-cognitive definitions of trust.

The last belief-condition of this definition (c) concerning trustworthiness will be examined in part III on degrees of trust; at this point I would like to make two remarks concerning the expectation in condition (b). First, I note that this expectation that S has about T *ϕ'ing* does not, on its own, *obligate* T to *ϕ*; in being trusted to *ϕ*, T does not incur a *duty* to *ϕ*.[4] We can see this by noting that, where T is a human agent, *ϕ'ing* might be an action that no agent can be obligated to do because it is a morally or legally impermissible action; for example, *ϕ* might be first degree murder. It is possible for someone, in trusting another, to expect the trusted to perform an impermissible action, and to believe that the trusted is trustworthy to do it, but it does not follow that the trusted is now morally or legally *obligated* to do that action (indeed, in such a case the trusted would seem to be obligated not to do the action he is being trusted to do). Clearly, no one would be thought morally or legally justified in doing a wrong based on the claim that he was being trusted, and thus expected, to do it; trust relationships cannot dominate moral norms. Also, T might be a non-human agent of which it is impossible to have

obligations, say a smart-machine such as a self-driving car or a surgical robot, or an animal such as a guide dog. In this case it is possible for S, in trusting a non-human agent to ϕ, to expect this agent to ϕ, but clearly impossible for the agent to possess any obligation or duty to ϕ, for even though smart-machines or animals can be (to an important degree) autonomous agents, and perhaps possess (to a degree) moral or legal rights, they are not *moral* agents to the high degree that is required to take on obligations and duties. A third reason why the expectation in trust cannot, on its own, obligate the trusted to do an action is that, in being trusted to ϕ, T might not know she is being trusted to ϕ; S might, in trusting T to ϕ, expect T to ϕ and believe T is trustworthy in this regard, but keep this trust a secret (perhaps S believes that informing T that she is being trusted to ϕ will increase the likelihood that T will fail to ϕ). However, a plausibly necessary condition for obligation is that the agent who is duty-bound to do something knows *that* she has this obligation and knows *what* she is obligated to do.

Second, the expectation in condition (b) is not simply a factual anticipation or prediction on S's part that T will ϕ (though these might accompany this expectation); it is *normative* – evaluative – in (at least) two respects, neither of which however is moral.[5] (i) In keeping with the conceptual content of trust, this expectation – and thus, the trust as well as the trustor – is subject to being evaluated by norms of *reasonableness*. So, for example, S *should* not, in trusting T to ϕ, expect T to do an action that is not practically possible for T to do, an action that is, say, beyond T's physical ability or an action that is practically (if not logically) impossible to do; that would be unreasonable. And S's expectation *should* be coherent, in the sense of self-consistent. For example, where ϕ_1 and ϕ_2 are necessary parts of ϕ'ing, it would be unreasonable, in trusting T to ϕ, to: (expect T to ϕ_1) and (expect T to ϕ_2), but not expect T to (ϕ_1 and $\phi 2$); this would be an incoherent expectation (resulting in an incoherent instance of trust). But such norms of reasonableness and coherence do not imply "moral norms"; it seems perfectly possible that, in trusting T to ϕ, S's expectation is reasonable (by norms of practical rationality), and yet it remains an open question whether this expectation is morally right or wrong and whether T's ϕ'ing is morally a good or bad thing.

(ii) In trusting T to ϕ, the expectation S has that T ϕ's is also normative in the sense that it is *about* an action that is subject to being evaluated by norms of *performance*. So, S expects that T *should* ϕ and not perform some other action instead, that T *should* ϕ completely and not leave the action unfinished, and that T *should* ϕ well or at least sufficiently to produce the desired result and not perform it poorly or insufficiently. Such incomplete, under-performance, or mis-performance would not only be described factually as leaving S's expectation unsatisfied and S perhaps materially at a loss; it would, importantly, *justify* S to feel *disappointed*, *offended* or *put-out* – *resentment* in Baier's analysis (1986) – toward an agent (T) that *should* have functioned or acted differently. Where T is a human agent, a normative expectation makes it possible for S to *blame* T or hold T *responsible* for letting S down; where T is a non-human agent

(say, a service animal) this normativity makes it possible for S to be *owed an explanation* or to *deserve corrective changes* in T's functioning (say, by retraining). Such norms of performance, however, do not imply that they are "moral norms"; for example, if ϕ is a morally wrong action, it is clearly still possible to evaluate ϕ by relevant norms of performance as having been done completely and well, and as fulfilling S's trust. But, to repeat the above argument, a normative expectation is something about S's (the trustor's) mental content; it is not an obligation T (the trustee) incurs by being trusted. That S *normatively expects* T to ϕ is one thing; that T *should* ϕ is quite something else.

If this understanding of the two-fold *non-moral normativity* of the expectation involved in trust is correct, then the ethics of trust requires the moral evaluation of a trust relation as a separate and an independent undertaking, not something already achieved simply by establishing a trust relation (as would be the case if every trust relation were inherently a morally good thing).[6]

2.2 The valuing relation

The relationship that holds between the trustor (S) and the action (ϕ) that the trusted (T) is being trusted to do is one of ***valuing*** – S in some sense values ϕ (or values T's ϕ'ing). It is natural to assume that this value is instrumental; S values ϕ because ϕ'ing will produce a consequence that S has reason or desire to bring about, and instead of S himself ϕ'ing – for whatever reasons – S trusts T to ϕ. To illustrate: Jack trusts Jill to take care of Jack's house while Jack is away on vacation; Jack desires that nothing bad happens to his house while he is away and Jack values Jill's house-sitting actions because he believes that they will contribute to this consequence. While the action ϕ may have value independent of the trust complex in which S values it, the value ϕ has that I am describing is relative to a trust complex; outside a given trust complex, ϕ might have no value for S. Related to this point, the value ϕ has for S is subjective in origin: ϕ has value (for S) *because* S subjectively values ϕ, not the other way around. It might be too strong to claim that we have in this relation of valuing between S and ϕ the general motivation an agent has to establish a trust relation, but it is clear that this value-relation is a necessary condition for trust; an agent who does not value ϕ, and much more so an agent who disvalues ϕ, will not trust any agent to ϕ.

Now there might be instances of trust in which ϕ's value to S is so small that S, in trusting T to ϕ, is not in any way harmed or damaged or disappointed should T not ϕ, but I will put this possibility aside and focus in what follows on trust complexes in which the stakes are relatively high for S. High-stakes trust means that S places high value on ϕ (or on T's ϕ'ing). This, in turn, gives us a trust complex in which S, to achieve what S values, depends on T to ϕ, and depends a great deal the higher the stakes; the trustor is *dependent* on the trusted because the trustor

values the action the trusted is expected to do.[7] Should T fail to ϕ, or should T betray S's trust, S is not only subject to disappointment, S might also be harmed, damaged, or put at a loss proportional to how important ϕ'ing is to S. Here we have, I submit, the reason why trusting others renders the trustor *vulnerable* and why there is a (widely acknowledged) connection between trust and *risk*;[8] if ϕ'ing has no value for S, then S risks nothing in trusting T to ϕ and it is hard to see how S becomes vulnerable to the possibility that T never ϕ's.

2.3 The executing relation

The relationship that holds between T and ϕ is one of ***executing***, *doing, performing* or *accomplishing* – T is to *do* that action and, thus, functions as a practical agent within the trust-complex. The symbol "ϕ" is meant to range over actions in a broad sense. It includes relatively simple acts that an individual (human or artificial) agent might do such as lending money or driving someone to a destination as well as relatively complex (sequences of) acts that require various degrees of cooperation on the part of several agents such as earning a college degree or landing a plane in difficult weather conditions or performing a difficult surgery; it includes large domains of action requiring the coordination of systems and networks of agents such as an international airline flight, or maintaining a country's national security, or returning a human to earth from space. It includes actions done by single human and single non-human agents (e.g. your friend, your friend's seeing-eye dog, your GPS device), as well as organized collections of human and non-human agents (e.g. teams, corporations, institutions, smart-grids and computer networks). And it includes the "negative" actions of refraining and un-doing as well as that of doing. The symbol "ϕ," then, is meant to refer to any given action it is possible for a human agent to trust another agent to do.

If T is being trusted (by S) to ϕ, *when* is T to perform ϕ? It is reasonable to assume that the conceptual content of trust often includes time-sensitive meaning such that if T is trusted to ϕ at or by a certain time and T ϕ's but not at the right time, then T has not fulfilled S's trust. For example, if Jill trusts Jack today to have deposited needed money yesterday into their joint checking bank account, then Jack lets Jill's trust down (today) if Jack doesn't deposit the money until tomorrow. The symbol "ϕ" is intended to capture this time-sensitivity; it refers to any given action it is possible for a human agent to trust another agent to *have done*, to *be doing*, or to *do at some future point*.

I want to argue for three necessary conditions on the relation that obtains between T and ϕ. The first two concern conditions that make it possible for T to be trusted to ϕ, and the third concerns a condition that makes it possible for T to fulfil an agent's trust by ϕ'ing.

First, for T to ϕ (and thus for T to be trusted to ϕ), *ϕ'ing must be a possibility for T* in the two-fold sense that, when the time comes for T to ϕ, T is able (i.e. sufficient with respect to the capability and effort it takes) to ϕ, and T's circumstances don't present major obstacles or interferences to T's ϕ'ing. If either of these weren't the case, then it would disqualify ϕ'ing as a possibility for T; clearly, T can't be trusted to ϕ if ϕ'ing is not, in this sense, a possibility for T. (I.e. it would be unreasonable (see section 2.1(ii) above) to expect T to ϕ, if T can't ϕ.) In the case where T is a human agent, this possibility is typically described as an *option*; for T to ϕ, ϕ'ing must be an option for T. In the case where the trusted is a non-human agent, however, "option" might be considered an odd term to apply and the term "possibility" seems more applicable.

The second necessary condition on the relation between T and ϕ concerns T's ability *not* to ϕ. Along with the possibility to ϕ T must have the possibility *not* to ϕ, for if T must ϕ (i.e. if it were not possible for T not to ϕ) then no one would have to *trust* T to ϕ – it would be certain that T would ϕ in the relevant circumstances whether anyone wanted this to happen or not. The ability not to ϕ represents T's degree of *autonomy* within the basic unit of trust; if it is not possible for T not to ϕ, then T is not autonomous with respect to ϕ'ing. In the case where T is a human agent, we might describe this condition by saying that not ϕ'ing must be an option for T and that this represents a necessary minimum degree of T's freedom within the basic unit of trust; surely, there is no reason to *trust* T to ϕ if, when the time comes, T will be coerced to ϕ. It follow that a minimal degree of the trusted's autonomy is a necessary condition for there to be trust. It is important to note that this trust-related autonomy, i.e. T's ability not to ϕ (when trusted to ϕ) means that T must have some incentive, motivation, inducement, or responsiveness to a situation to act in the direction of ϕ'ing rather than in the direction of not ϕ'ing.

In the case where T is a non-human agent, I will distinguish (in part III of this project) between T's autonomy and a breakdown in T's functioning. To illustrate: suppose a plane's human pilots (S) activates the plane's automatic pilot (T) and trusts it to land their plane (safely, of course) in foggy conditions (ϕ). If T has no autonomy, then T will land the plane even if landing conditions require the landing attempt to be aborted (not ϕ'ing); the ability to abort an overly risky landing must be an "option" for the plane's automatic pilot or it couldn't (or shouldn't!) be trusted to ϕ. However, if the plane's automatic pilot software crashes such that there is a breakdown in it functioning, then it is also the case that aborting an overly risky landing is not a possibility for the plane's automatic pilot, and it can't (shouldn't) be trusted to ϕ; but in this case we are not dealing with an "option."

The third condition on the relation between T and ϕ concerns the possibility of T satisfying or fulfilling S's trust by ϕ'ing, where S trusts T to ϕ. The condition is this: *The description under*

which T performs ϕ should match the description under which it is being expected by S. To illustrate: suppose Jack (S) trusts Jill (T) to *lend* Jack a certain amount of money (*ϕ*), and Jill does not *lend* Jack the money but instead makes it a *gift* to Jack. We have a transfer of the money from Jill to Jack, but from the perspective of Jack's trust, Jack's expectation is that the money be a *loan* whereas from the perspective of Jill's action the money is a *gift*. It is clear that these descriptions under which a single transfer of money are "understood" (*as* a loan verses *as* a gift) don't match; they describe quite different actions even though the movement of money on Jill's part to Jack in each case is the same. So, we have a case in which Jack trusts Jill to *ϕ* and Jill does not *ϕ* but does something else instead, the result of which, however, is the same: Jack receives from Jill the money he needs. We must conclude, I want to argue, that Jill does not fulfill Jack's trust, and that Jack would be justified to consider his trust let down or disappointed by Jill, because – even though Jill does the behavior *that* Jack, in trusting Jill, expected Jill to do (i.e. give Jack the money) – Jill does not do it *as* Jack expected her to do it (i.e. as a loan, rather than as a gift); the description under which the trusted does the action does not match the description under which the trustor expected it to be done in this trust relation. So, a necessary condition on the relation between T and *ϕ*, if T is to satisfy or fulfill the trust T is given to *ϕ*, is that T performs *ϕ* under the same description the trustor expects T to the action. If this description from the perspective of S's trust is more general, broad, or loose, then T is able to satisfy S's trust by a wider range of possible actions; if the description from the trustor's end of the trust relation is more specified, then T is more constrained in performing the action that fulfills S's trust.

I will now turn to the problem of kinds, degrees and forms of trust – as these were presented in section 1 – and in what follows specifically to kinds of trust, using the basic unit of trust presented in section 2 as my framework. This basic unit, we recall, has a 3-part form: (*S trusts T to ϕ*). A relation of *trust* links a human trustor (S) to a (human or non-human) trustee (T), a relation of *valuing* links a human trustor (S) to an action (*ϕ*), and a relation of *executing* or *performing* links a trustee (T) to the same action (*ϕ*) that the trustor values. The trustor, in establishing a relation of trust with T (to *ϕ*), intentionally (or decisionally) trusts T, expecting T to *ϕ* based on the belief that T is sufficiently trustworthy to *ϕ*.[9] We see that this basic unit is *practical* trust, i.e. trust involving practical agents concerning actions and their utility; it is not *epistemic* trust, i.e. trust involving epistemic agents concerning information and its (probable) truth.

3. Kinds of trust

Our experience of trust makes us aware that we trust a variety of human and non-human agents, but we don't trust them all in the same way. In the literature on trust, trusting a single human agent – *interpersonal* trust – is taken as the standard case for exploring trust, and insights from this standard case are then applied to "higher levels" of potential trustees: organized collections of human agents (e.g. groups, institutions, corporations, levels of government). On this basis authors have examined, for example, institutional trust, economic trust, social trust, and political trust.[10] When trusted to ϕ, the required level of organization and ability of such aggregates of human agents to ϕ are typically mediated by, assisted by, and enhanced by digital technology (e.g. communication systems, software programs and the hardware that implements them, smart power-grids, AI systems, robots, and a variety of "smart-devices"); as such, there is a sense in which trusting these kinds of collections of human agents extends our trust to a variety of multi-agent systems that contain both human and artificial agents.[11] We see that the general strategy in this effort to understand different kinds of trust is to start with different categories of agents that might receive our trust, different kinds of potential trustees, and then distinguish kinds of trust and ways we trust in line with these categories; possible trustees lead to possible kinds of trust. The challenge faced by those who work within this strategy is to discover whether, or to explain how, it is possible for certain kinds of non-human agents to be the receivers of our trust relative to how much they deviate from the standard model of trust as the starting point: one-way interpersonal trust between single human agents.

My plan in this section is to explore an alternative strategy; one I believe is closer to our experience of trust than the above strategy. I want first to ask: on what basis do human agents trust – what are the basic ways a trust complex happens? The ways that a trust complex comes about and is sustained will give us three kinds of trust. I will then try to fit various kinds of agents as potential receivers of trust into these kinds of trust. I believe that the insights won by the standard strategy used by trust researchers won't be lost; they can be nicely included within the alternative scheme I propose.

Given the basic 3-termed trust complex: *S trusts T to* ϕ, I want to distinguish three ways this complex comes about: (i) communication-based trust, (ii) practice-based trust, and (iii) context-based trust.

3.1 Communication-based trust

Communication-based trust is any trust complex in which S and T form a trust relation by communication.[12] This is perhaps the most familiar, though not the most occurring, kind of trust. Either S informs T that she is (or will be) trusting him to ϕ, or T requests S to trust him to ϕ ("Trust me in this."). This communication might happen by language use or by some exchange of outward signs and signals between S and T; it might be direct or communication mediated by a go-between. In any case "communication" means that T tells (or indicates to) S that T understands and agrees to do what S tells (or indicates to) T she desires or expects T to do. The trust relation between S and T is understood by each, even if the communication has not explicitly referred to "trust" by name; this is to say, by such communication the unit (*S trusts T to* ϕ) becomes *common knowledge* between trustor and trustee. Here is a typical illustration: Jill trusts her friend Jack to take care of her house while she is away on vacation. One possibility is that Jill *asked* Jack to do her this favor and Jack, understanding what Jill is requesting, *agrees* to do Jill this favor. Another possibility is that Jack *asked* Jill to trust him to take care of her house while she is away, and Jill *agrees* to let Jack do her this favor. Either way, communication between Jack and Jill not only gives rise to this trust complex (that Jill trusts Jack to take care of her house while she is away in vacation), it means that each has agreed to be (an important) part of it as trustor or as trustee, and that each knows that both have so agreed, and each knows that both know that both have so agreed, … .

It is easy, I believe, to make too much of communication-based trust; its apparent strength has its problems – part of what Judith Simon calls "the dark side of trust."[13] For example, compared to the other kinds of trust (see below) communication-based trust provides perhaps the best opportunity for the trustee to deceive, exploit, or betray the trustor, because the verbal assurance that establishes this kind of trust complex, and the likely continued verbal re-assurance that would work to maintain it, could well lower the trustor's reasonable skepticism that any trust complex should contain. It is easy for the trustor to become *too trusting* in communication-based trust complexes. Also, this kind of trust has the potential to weaken the trustee's ability, after agreeing (committing) to do what the trustor expects (ϕ), not to ϕ should T start to have misgivings about ϕ'ing. In this respect, communication-based trust can not only lower or compromise T's sense of his/her autonomy, it can make T more vulnerable than necessary to being deceived or exploited by the trustor.

With regard to the last point, it is perhaps worth noting that communication-based trust, on its own, cannot obligate T to ϕ. Even if, in addition to their mutual understanding and agreement, T *promises* S to do what S trusts T to do, T is still not morally or legally obligated to ϕ. The reason for this is that no agent can be morally obligated to do an immoral action, and no agent can be legally obligated to do an illegal action, (though an agent can be morally obligated to do an illegal action and legally obligated to do an immoral action), and ϕ might be a morally or legally impermissible action. However, it is safe to say that T would have a (moral) duty to ϕ,

not because S trusts T to ϕ based on their communication, but because: (a) T promises S that T will ϕ, and (b) ϕ'ing is morally permissible.

I now turn to the question: what kinds of agents can enter as trustee into a communication-based trust complex? It appears to me that two types of agents qualify. Obviously, single human agents (HAs) form the first category that readily comes to mind (subject to the conditions stated above in sub-section 2.3). This yields the standard case of *interpersonal* trust. The less obvious category is that of single artificial agents (AAs) having language ability: *interagential* trust. While still relatively primitive, such "smart" AAs are increasingly gaining impressive communication ability, to the point that they are trusted to act in specific (albeit dedicated) ways in high stakes trust situations. Three examples will illustrate.

In the health care world, vital medication is now ordered by HAs via telephone discussions or internet exchanges with such AA's that then assure the trustor that the order has been understood and inform the trustor when to expect delivery. The HA clearly trusts the AA to have the correct medication sent within the stated time, and much could depend on the AA fulfilling the HA's trust. The AA provides all the verbal or written assurances and polite communication that would (or should!) take place between two HAs.

In the world of smart-phones, AA's having language ability are no longer a novelty, many smart-phone operating systems now include their own (typically "female") AA with which the phone's HA owner communicates. The interaction not only involves epistemic trust between epistemic agents (e.g. to be accurate in exchanging information), the communication gives rise to practical trust. So, the HA is able to communicate directives and requests to the AA, which responds with an acknowledgment of "understanding" or a request for clarification. On the basis of this communication, the HA trusts the AA to carry out such directives, especially when the actions are important to the trustor such as financial transactions, making appointments and communicating reminders, and purchasing and sending gifts.

In the world of smart-automobiles, not only the GPS has the ability to communicate with HAs, if recent projections are correct soon the entire vehicle will be a system of AAs that communicates with its (authorized) owner: recognizing the HA's face, handprint, and voice, processing verbal directives, and carrying out the transportation and entertainment activity the trustor expects. As we know from reports of both comical and serious mistakes on the part of GPS AAs, HAs clearly trust their GPS AAs to provide useful driving directions and traffic reports based on communication between trustor and trustee.

Aside from individual HAs forming communication-based trust relations with other individual HAs and with individual AAs, what about aggregates of HAs as potential trustees by communication, social and business institutions for example? I believe that these are best

understood as falling under other kinds of trust, with perhaps one exception: the case where such an organization has a single HA (or AA!) as its legitimate representative with which one can communicate. For example, I see no conceptual problem with trusting a government agency based on communication if the government is a dictatorship and the trustee with which one communicates is the dictator, for in this case there is no question that the government's representative "speaks" for the entire government.

3.2 Practice-based trust

Practice-based trust is any trust complex in which S and T form a trust relation based on a set of practices in which they engage.[14] Practice based trust is perhaps the most commonly occurring way that we form trust relations. Typically, no communication is necessary, and any communication that happens to take place is not sufficient to place the trust in the category of communication-based. A "practice," in this sense, is a publically recognized pattern or routine of behavior, governed by a set of more-or-less formal or by conventional (informal) norms, that is general in nature (i.e. does not have to be performed by a special individual). Examples of practices include: the convention of keeping to the right when going up or down crowded public stairways, voting in local or national elections, using public or commercial transportation systems, buying or selling items and services, the standard activities that make up your profession, career or job, and the use of the Internet. In this kind of trust, the trustor and the trustee are engaged in coordinated, often cooperative, practices and S trusts T to do his/her/its part according to norms and conventions of the practice(s) in question. A familiar example illustrates practice-based trust.

You are driving on a single lane road that has oncoming traffic. You and these other drivers are engaged in a publically recognized and reasonable understood practice, largely governed by traffic laws, but also by "rules of the road" conventions. You know that it can happen – and that it actually has happened to other drivers – that an oncoming vehicle swerves into your lane; and you know that the result can be, minimally, a costly traffic accident and, more likely, serious injury or death to you and any passengers in your car from such a head-on collision. Yet this possibility does not keep you from driving (assuming that you are not forced to do so). Instead, you trust each oncoming driver (either individually or collectively) to drive according to the norms of the practice in which you are both engaged, e.g. not to drive recklessly, not to fall asleep, not to become overly distracted in conversation or daydreams or with electronic devices, not to be intoxicated, not to give into a suicidal urge, etc.

Keeping to the world of driving, you know that it has happened to drivers stopped for a traffic light or for a stop sign that a rear driver fails to stop and causes a rear-end collision, resulting in serious head and neck injury. You know that this could happen to you every time you stop for a traffic light or stop sign and there are drivers behind you, yet this possibility does not make you take unusual precautions each time you approach a red traffic light or stop sign. Instead, you trust the drivers behind you to do their part of the practice in which they are engaged.

Throughout the day, people are engage in numerous practices and by such practices interact with each other in various capacities and roles with either no or minimal communication: professionally, socially, economically, and politically. The agents need not know one another personally or have direct contact; they can interact anonymously and "at a distance," as in the practice of mailing an important package to a family member, and the cooperative practice(s) of the postal system processing and delivering your mail. Many, if not most, of these interactions could not be sustained and would not continue without trust, and these trust relations are based on the fact that the agents enter and engage in their respective practices, and thereby take on the role of trustor or trustee.

What kinds of agents can enter as trustee into practice-based trust complexes? Clearly individual HAs qualify; but also domestic animals can function as trustee. By a combination of breeding and training, the practices of guarding valuables, detecting certain chemicals, search and rescue, herding, and guiding blind HAs are carried out by dogs. Again, primarily by breeding, both cats and dogs engage in patterns of behavior of hunting (and thereby the population control of) dangerous disease-carrying rodents, important health practices in many parts of the world. HAs can and do place a great deal of practice-based trust in their service animals to do their part in various high-stakes trust complexes.

Individual "smart" AAs also engage in a variety of practices that HAs trust them to perform. For example, think of: computer generated financial trades in national and international markets, automated tellers in banking transactions, automatic pilots that fly and land airplanes, and the various "smart" systems that control certain functions in your automobile, such as automatic breaking in slippery conditions or when certain front or rear objects are detected, stabilizing in potential roll-over situations, and driverless parking. Our digital technology, it appears, is increasingly providing HAs with AAs – smart-machines and intelligent systems – that engage in and sustain practices that HAs, by engaging in corresponding practices, trust them to do.

I believe that the largest category of trustee in practice-based trust complexes is groups: aggregates of HAs and AAs organized to function as a unified practical *agent* and that typically have an overarching purpose they work to achieve or service/product they offer. Such groups are themselves AAs; I will refer to them as GAAs (group artificial agents). GAAs engage in complex, often time-consuming, practices that no individual HA or AA can perform, requiring

the coordination of many sub-practices. We can think for example of educational institutions, local and national government agencies, and "industries" such as food production businesses, health-care providers, airline companies, and home construction services. Putting aside instances of communication-based trust that might take place between HAs and individual within such GAAs, by engaging in complex aggregate practices GAAs interact with individual HAs who function as receivers of the services/products that such GAAs provide; HAs are receivers of, for example, professional education and training, international airline flights, preventive and emergency health-care, food at restaurants, public transportation, newly built houses, and civic services such as highways cleared of ice and snow, utilities, and traffic control.

By engaging in practices that interact with those that GAAs engage in, HAs often form practice-based trust relations in which GAAs are trustees. We can see this in high-stakes cases in which a breakdown of a GAA's practice results in a significant worry, loss or injury to the trustor; say, a case of food poisoning at a restaurant, lost baggage in an airline trip, the cancelling of an important training program at a college one is attending, on-line identity theft, or error in the results of your blood-tests connected with a medical examination. We know that such breakdowns happen and when they do, the HA "victim" quickly realizes how much he or she trusted the GAA in question to perform its end of the practice, and how offended the trustor has every *right* to feel in having his/her trust "violated." Yet we still decide to engage in a variety of practices that interact with a variety of GAA practices; when we do we are (often) trusting GAAs to perform sufficiently the practices we (normatively) expect of them.

In sum, I have argued in this sub-section that in the unit (*S trusts T to ϕ*) where S is a HA, a practice-based trust relation is possible with T's that are: individual HAs, individual domestic animals, individual "smart" AAs, and GAAs containing HAs or HAs and "smart" AAs.

3.3 Context-based trust[15]

Context-based trust is any trust complex in which S forms a relation of trust with T as part of the broader framework or context in which S lives. "Context" is not a precise concept. Intuitively, it means the community of which S is a member, in which S is "officially" accepted and perhaps feels a sense of belonging, and from which S gains an identity (as a member of such-and-such community). As I mean it, "context" also means something like a "life" in the sense of a pervasive quality or profile of the kind of life one lives, or a "world" in which a person spends significant amounts of time. The idea is that a person typically lives several "lives," or lives in several "worlds," or belongs to several "communities." For example, a person professionally living the "life of an academic" might also be part of a family living a "farm life." "Context" in this sense is close to Wittgenstein's influential concept of a "form of life", i.e. the

normative system of agreements and acceptances in which a person works, plays, pursues important (short or long-term) projects and interests, realizes potentials, and by which his or her life has meaning and is, at least in part, "formed" and "justified" and has "its value."[16] The "life of a retiree" or living a "blue-collar life" or the "world of a professional musician" provide further examples of "context" as I intend to use the term; these name rich, complex human spaces or environments or communities containing networks of people, familiar conventions, routines and practices, available (if not always welcoming) institutions, and a degree of geographical and normative stability.

Now in one sense, every instance of practical trust is "contextual"; after all, trust takes place at a particular (span of) time, in a particular situation containing particular people who are doing certain activities. A "contextual theory of trust," thus, might mean a theory that argues that, for all trust, the context in which it occurs is essential to it. However, "context-based trust," as I mean it, is not such a theory. I am proposing that one of the ways that the trust relation is formed, how a certain kind of trust comes about, is based in the trustor's "context." Context, I will argue, is (or provides) the *mechanism* by which HAs are able to trust "remote" agents – agents with which the trustor has no direct interaction or proximate contact either temporally or spatially – and trust larger "systems." It seems to me, for example, that there is a very real sense in which we now trust past HAs who are no longer alive (e.g. past authors of a document that founds a current government or guides a current way of understanding human affairs) as well future HAs who are not yet alive (e.g. future generations to care for endangered species, or to continue a multi-generational project), and that we now trust contemporary HA and AAs with which we never have had and never will communicate or directly interact with by engaging in a practice.[17] The question is: how is this possible? How best to understand this kind of trust and its reach? I want to argue that a trustor's *context* makes this possible and gives us a way to understand this kind of trust.

A clear, even if a-typical, example of a context (in this sense) will help me explore context-based trust: the International Space Station (ISS). This example has the virtue of being a relatively isolated "idealized world" in which an HA lives a certain life – the life of an astronaut; it is a context that gives its inhabitants clear (functional) identities, assigns them familiar routines, provides a feeling of (technological) security, and one that contains systems and networks of HAs and smart-AAs with which a given HA interacts. Of course, the ISS is not the typical earthly context; aside from its human inhabitants, the natural environment (on the macro-level) that one ordinarily finds in earthly communities is absent; in addition, the ISS is subject to a degree of design and control one would not typically find in the "loose" and "haphazard" organization of the earthly "worlds" in which we live our diverse "lives." Nevertheless, it shares with earthly contexts certain structures and features that lets it serve, for my purposes, as a relatively

"pure" illustration of a "world" or a "life" – i.e., a context – by which to explore context-based trust.

I don't see how anyone could doubt that an astronaut (Sally) living in the "world" of the ISS is an actively trusting HA and that this trust has a far and wide reach, and is "global" in quality in so far as Sally trust, in some sense, the whole ISS and the systems on which it depends. Some of the agents Sally trusts are extremely remote such that it could not be trust by communication or trust by practice. I don't mean to deny that an astronaut living in the ISS won't or can form particular communication-based and practice-based trust relations; to the contrary, it is hard to see how these could be avoided in such an environment. But in addition to such proximate trust relations, it is reasonable to think that Sally also trusts remote agents to perform various important actions. For example, there are HAs and AAs on which the functioning of certain systems depend, on which the functioning of other systems depend, … , to provide Sally with water, and that Sally's mental state contains the three elements (decision, normative expectation, belief) for establishing a trust relation with such agents to provide her, via a complex sequence of processes and systems, with water. It is conceivable that such a chain of dependencies goes back to HAs who are no longer professionally active or even alive, or to AAs that no longer operate.

There are two ways of thinking about Sally's trust of such remote agents and her global trust of her context that I believe are wrong: a reductive way and a holistic way. It might be argued reductively that Sally's trust in such cases is no more than an *aggregate* of different communication-based and practice-based trust relations, and that (generalizing) any trustor's "trust" of remote (temporally or spatially distant) agents, sequences of processes and whole "worlds" is best understood as always reducing to individual trust relations of particular HAs and AAs whose actions maintain their workings. I see two problems with this position: first, it misses the obvious fact that the ISS functions as a *whole* system to support Sally's life and provide her a reasonably livable world; it is, in an important way, a unified "world." Sally – once comfortable with her life within the ISS – trusts this whole "community" which sustains her, without necessarily forming an individual trust relation (or even knowing about) each and every HA, AA, and system that has a role in or contributes to this context. Sally's trust, it would seem, and by extension a certain kind of human trust in general, can have the quality of being "global" or "general" – perhaps "holistic"—in extending to or reaching throughout her entire "world" without performing and aggregating all the mental acts of trust that are possible with respect to each and every agent involved in the ISS.

Second, it is reasonable to think that the human capacity for trust is not only finite, it is relatively limited. As a human mind can think of only so many individual numbers at the same time (or in a given unit of time), by analogy a human trustor can trust only so many agents;

human trust is not infinite. The ISS is a massively complex "world," such that it is would appear beyond Sally's capacity to trust to establish distinct individual trust relations with each and every agent that contributes, proximately or remotely, to its functioning.

What about the holistic position? Why not say: Sally (S) trusts the whole ISS (T) to function as designed (ϕ). One problem here is that the ISS is not itself an *agent*, it is a context in which agents live certain lives; thus, given our definition of trust as a relation between agents, it can't be a trustee. GAAs such as corporations, institutions, and government agencies are organized as, on some level, single unified *agents*, and as such can function (as we saw above) as trustees in a trust relation. While such GAAs are important parts of Sally's context, they are not themselves "communities" or "worlds." A context such as the "world" of the ISS, or a "life" such as that of an astronaut in this world, is a model for the complex earthly "worlds" HAs typically live in, and these are hardly single unified agents.

Another problem with the holistic approach is that it seems to commit the fallacy of division. It would be fallacious to argue from whole to parts: (1) The world in which a trustor lives is a whole containing a variety of interconnected HAs, AAs, processes, and systems; (2) The trustor trusts her whole world, it is the trustee; (3) therefore, the trustors trust extends to all parts of her world, they are all trustees. Such an argument, aside from being fallacious, leaves it unexplained *how* Sally's trust propagates throughout the ISS, and *how* it is able to bypasses its parts (for example, her electric toothbrush) that Sally might not include in her trust.

How, then, should Sally's (and our) remote or global trust be understood, if not reductively or holistically? The ISS is, by design, a *network* of relations and interconnections of a variety of agents, processes, and systems. Three interconnections are important for context-based trust: dependency, reliance, and causality. (i) There are relations of *dependence*, in the sense of necessary condition; ISS system X *depends on* ISS system Y iff X cannot operate/function unless Y operates/functions. For example, if the ISS lighting system *uses* solar panels as its source of electrical power, then the operation of the lighting system *depends* on the operation of the solar panels in the sense that if the solar panels don't function, then the ISS lighting system does not function. (ii) I will call it a relation of "*reliance*" when a HA *depends* on some system in the ISS. For example, Sally *relies* on a computer (she has named "Hal") to monitor her sleep and wake her only after a certain amount of sufficiently deep sleep. (iii) There are relations of *causality*, in the sense of sufficient condition; the operation/functioning of ISS system X makes (produces, causes) ISS system Y to operate/function – X is cause and Y is effect. For example, a piece of software on the ISS is activated *when* sunlight is detected by a sensing instrument on board; these two ISS systems are connected by a cause-effect relation.

The ISS is a network of interacting agents, processes, and systems based (in large part) on these relations of dependence, reliance, and causality; different parts of the ISS can "talk" to each

other and effect each other only to the degree that they are connected by one or more of these three relations. The connections between any two parts of the ISS need not be direct; dependence, reliance, and causality form chains of single or mixed relations linking each part of the ISS with every other part; and these relations are (widely believed to be) *transitive* and, thus, the ISS's network possesses a high degree of interconnectedness. As a result, the ISS is not only an internally systematic and coherent "space," it is also a world that extends far and wide. It reaches, by these chains, to a variety of more-or-less spatially distant human and artificial agents, support and communication systems, and smart-devices of space technology, located in other satellites orbiting earth, on earth, or perhaps on other planets (e.g. the Mars rover). These chains of causality, reliance, and dependence not only operate in Sally's temporal present, they extend to past and to future agents, processes, and systems; the ISS, as with every human context, is affected by its past operations and by the goals it is designed to realize at various future points; that is, by its history and by its future. This, then, is the picture of Sally's context I want to present and use as a model for defining context-based trust.

I will assume that there are direct (proximate) relations of *trust* by communication and by practice that Sally forms with a variety of other HAs and AAs in the ISS and on earth. Because the trust relation is not transitive, it can't on its own form chains; it can't on its own "spread" from Sally to remote systems and agents, or "permeate" her whole ISS world. But the relations of dependence, reliance, and causality, we see, do form a vast and complex network of such chains. My suggestion is that Sally's initial communication and practice based trust propagates throughout her context by "riding" on the network it contains. That is, remote trust relations between S and a variety of HA and AA trustees, and eventually global trust between S and S's contexts, are possible by being "carried," or "supported" by – i.e., by *supervening* on – chains of dependence, reliance, and causality that link various agents and systems with S's context. This context-based way that trust relations become established can be stated in terms of principle CT:

(CT) Within S's context, if S trusts agent T by communication or by practice to ϕ, and T is linked to X with respect to ϕ'ing by a network chain, then S's trust extends to X *by context* with respect to ϕ'ing.

The intuitions behind trust by context are:

(i) if you trust the effect, then your trust extends to the cause; and
(ii) your trust extends to whatever is necessary for what you trust.

That is, relations of causality and dependency can extend your trust beyond its immediate and direct target, but the extended trust is not the same kind as the initial trust. Of course, putting the idea this simply opens it up to counterexamples that a correct understanding would

exclude. So, I am not claiming, for example, that: to trust the son to ϕ is to trust his parents, for they are his cause. Nor am I claiming, for example, that: to trust the son to ϕ is to trust the element carbon, for carbon is physically necessary to the son's ϕ'ing. I am suggesting, however, that if the son can't ϕ without depending on his parents (with respect to ϕ'ing), then trusting the son (say by communication) to ϕ means that the trustor's trust extends to the parents *by context* not to make their son's ϕ'ing impossible; and if the parents can't do this without depending on … . In this way, context-based trust is "founded" on and "tracks" the various chains and links that give our earthly contexts their network structure (*relative to ϕ'ing*), and by this mechanism a trustor's trust gets to "permeate" the trustor's context. But, importantly, context-based trust can't reach any part or agent of the trustor's context that is not a dependency link, or a causal factor, in its network; they wouldn't be trusted anyway (or would have to be independently trusted). It would be an implausible theory of context-based trust that required us to say that Sally's trust extends to, say, her toothbrush to work as designed.

I define context-based trust between practical agents formally as:

> S trusts T by context to ϕ iff:
> (i) T is temporally or spatially remote from S,
> (ii) T is an agent within S's context (C),
> (iii) S trusts human or artificial agent U by communication or by practice to ϕ, and
> (iv) agents U and T are linked by a chain of dependency, reliance or causality in (C) relative to ϕ'ing.

We see, if this analysis is correct, that the context in which S "lives" provides the mechanism (the subvenient or founding base) for an HA to trust temporally and spatially remote agents and to trust them in a way that has a "global" reach; yet context-based trust is a kind of trust that remains within the basic 3-part unit: (S trusts T to ϕ).

I want to note that context-based trust is not reducible to particular instances of other kinds of trust, in part because it has a unique feature that neither communication- nor practice-based trust has: the context in which this kind of trust takes place *contains* the trustor as a part, and contains the trustor's trust as contributing to its network structure. The ISS, to return to our model, is the context in which Sally "lives" her life, and in doing so she makes a significant contribution to her world, helping to sustain it, playing a (more-or-less) important role in making it an acceptable and accepting human "space." As a node in its network, chains of causality, dependency, and reliance lead back to Sally as well as away from her. Thus, Sally is, indirectly, her own *remote trustee*, and her "global" trust of the ISS to work as it is designed includes her and her context based trust in its reach. Context-based trust, then, contains an element of (indirect) self-trust – the trustor also wears the hat of trustee, and as such is subject to norms of trustworthiness (to be examined in part II of this project). This is not the case with

either communication-based trust or practice-based trust; each of these kinds of trust is compatible with non-self-trust, and even self-distrust, on the part of the trustor, but context-based trust is not.

3.3.1 Comparison of context-based trust with the Buechner-Tavani contextual theory of trust

I believe that this concept of content-based trust gains significant support from, as well as offers support to, the work done in this area of the philosophy of trust by Jeff Buechner and Herman Tavani. Buechner and Tavani (2011) and Tavani (2014 forthcoming) have developed an insightful "contextual theory of trust." They build on Margaret Urban Walker's work on trust as a default attitude or outlook agents have in certain "zones" or "communities" in which they live in relative "ease, comfort, or complacency that relies on the good or tolerable behavior of others" (Tavani and Buechner 2013, Tavani 2014 forthcoming). This kind of trust is the default, in the sense that it is the context-based way an agent automatically and naturally lives and functions in a given zone; that is, until something goes wrong and the trustor's normative expectations are disappointed. In their model, such zones of default trust (or at least some of them) are not directed to any particular agent or focused on any given practice or institution, nor is it directed to a whole "zone" as a unified single agent: a "community" as a single trustee as it were; rather they develop Walker's idea that (some) zones of default trust are "diffuse" in the sense that default trust can spread throughout the entire "zone" or "community," to all its interconnected networks and systems of human and "smart" artificial agents.

Tavani (2014 forthcoming) and Buechner, Simon, and Tavani (2013) describe the kinds of agents that might populate such zones of diffuse, default trust (DDT); their aim is to explain how and to what degree AAs can be receivers of our trust (which they define as a disposition involving a normative expectation), and their argument is that certain AAs can enter a trust relation (as trustee) if the trustor's trust is DDT, and the AAs in question "live" in the zone receiving this trust.[18] These authors argue for different "strengths" or "levels" of DDT depending on such factors as: the level of "sophistication" an AA possesses (they distinguish four levels based on James Moor's four-fold classification of agents having ethical impact), how directly or indirectly S interacts with different AAs within a given zone, how much a trustor risks or is vulnerable should different AAs fail to meet the trustor's expectations, and how functionally autonomous a trustee AA is.

There are clear similarities and mutually supporting insights in the concepts of context-based trust (as described above) and Buechner-Tavani (B-T) model of zones of DDT, especially with regard to trust being "global" or "diffuse" with respect to context. However, one significant

difference stands out. Trust, in the B-T model of zones of DDT, is a *default* disposition. If my understanding of their position is correct, trust is a default disposition that automatically and "naturally" develops within the trustor as s/he comfortably interacts with familiar agents, systems, and institutions in a given "safe" zone of activity. The trustor, in this model, becomes aware *that* s/he has such a (presumably "silently operating") trusting disposition only when something goes wrong and disappoints or lets the trustor down and the trustor has resentment, and perhaps distrust, not toward any particular agent(s) or "generic" function(s) within the zone that was singled out for trust, but toward the general "workings" of the zone. Such DDT, as these authors present it, is a disposition that seems akin to what we ordinarily referred to as "taking for granted" as in: we take our beating heart for granted, until something goes wrong with it. DDT, on this view, seems to be a kind of "thoughtless trust," more an absence of worry or lack of concern or complacency than a consciously formed and sustained mental state, the product of a decision.

There are three problems with the model of trust as a default disposition within a zone that the model of context-based trust does not have: one is factual, and two are conceptual. Think of the factual problem as Hobbesian: why is trust and not distrust the default disposition? Empirically, it would seem that most humans most of the time live their lives in zones of relative danger, threat, oppression, insecurity, struggle, and discomfort. This includes minorities who live in prejudiced or unwelcoming communities, women world-wide who are subject to a variety of inequality and violence, societies that experience on-going low-grade ethnic or religious hostilities, individuals who live in high-crime neighborhoods, the poor whose lives are challenged daily to make ends meet, and the exploited who must practice self-protective skills and can't let their guard down. How does the theory of zones of DDT address this factual challenge: aren't areas of human life in which trust is the default disposition extremely rare, enjoyed – if they exist at all – perhaps only by the well-off, those living comfortable and secure – that is, privileged – lives? Isn't it the case that the real "worlds" in which most humans live, at best, call for both trust and distrust as agents and situations change, which they constantly do? Aren't zones in which "thoughtless trust" is the *rational* default too ideal and rare to serve as a model of trust within actual human communities? It would seem that the occasions in which human agents are happily surprised that things work in the zones they occupy outnumber (perhaps by far) the times in which human agents are disappointed/resentful that things don't work. And it would seem that the relatively lucky, easy, and "blessed" lives of a minority of agents for whom trust, rather than distrust, is the default in their first-world communities is not the appropriate model for the bulk of humanity; doesn't it go against what we have come to accept as the "human condition," and against the understanding of human nature and development within evolution theory? At any rate, an answer to this challenge is more an empirical than a philosophical matter.

One conceptual problem with the model of zones of DDT is with the concept "default." A default position, by definition, is one that needs no explanation or justification; the "burden" is to explain or justify or require evidence for *departures* from the default. The authors, however, justify DDT as a rationally appropriate disposition and thereby undermine its default status. To illustrate this point, consider a light with two possible states: either on or off. The light changes its state depending on a motion sensor. If the off state is the default, then an explanation is required why the light is on; there is no requirement to explain why the light is off, for "default" already exempts the off state as needing to be justified. To explain *why* the off state is the default, one simply points out: because we want the light's on state to indicate motion. That is, it makes sense to say that the light is on because the motion sensor detected motion in its range of sensitivity; but it is incoherent to say both that the off state is the default and that the light is off because the motion sensor is detecting *no* motion in its range of sensitivity, for such an explanation makes the off state not a default position but one of equal status to the on state. What should be said is that the light's off state is its default position and that the motion sensor is *not* functioning, i.e., *not* detecting motion in its range of sensitivity.[19] By analogy, making trust the default disposition in an agent's zone of activity means that this trust requires no explanation or justification, that it is the "un-designed" – the natural, *non-normative*, automatic – state of an agent's zone of activity. Yet, within the B-T theory of zones of DDT, it is argued that such default trust is justified, i.e. explained as the agent's *normatively* appropriate response to his/her safe and welcoming environment (until sometime goes wrong and trips the response of distrust or resentment). Conceptually, explaining/justifying a kind of trust as the normatively appropriate disposition in response to a zone of "ease and comfort" makes this kind of trust *not* a default position.

The second conceptual problem with making trust a default disposition in an agent's zone of activity is that the trusting agent cannot *decide* to make it the default disposition. As soon as the agent *decides* to trust, it is not a non-normative, automatic "un-designed" default position. That is, as a default, this trust cannot be a mental state that the trusting agent intentionally activates as a reasonable response to the agent's environment. Default trust within a zone, it would seem, is an involuntary state that automatically *happens to* the trusting agent simply by inhabiting the zone with "ease, comfort, or complacency that relies on the good or tolerable behavior of others." Yet this trust, by definition, includes the element of normative expectation that what the agent is familiar with and is used to occurring, will continue to occur in familiar ways. But, it would seem, here we have a problem: is it possible for a normative expectation to take place in an agent in the way a default disposition takes place, i.e., involuntarily and "thoughtlessly"? Clearly not; an "*involuntary, thoughtless, non-normative, normative expectation*" appears to be an impossible mental state-of-mind or achievement. The problem is that to the degree that trust involves a *normative expectation* it cannot arise as a default

disposition, and to the degree that trust is a (non-normative) default disposition it cannot be a normative expectation.

In contrast, context-based trust is not a default state-of-mind, and so does not have the above factual or either of these conceptual problems; it is trust that remains within the definition of a mental state that combines three mental *acts*: a decision, a normative expectation, and a belief. Aside from this (important) difference, however, it is clear that the above model of context-based trust and the B-T model of zones of DDT have many overlapping insights.

4. Concluding remarks

I began this paper with the claim that in our experience of trusting and being trusted we become aware of three central features of human trust: there are *degrees of trust* that makes it possible to trust unreasonably (i.e. excessively or deficiently) or reasonably; there are different *kinds of trust* relations we form and maintain with a variety of agents; there are different *forms of trust* that can take place with different kinds of agents. Failing to distinguish these three features leaves a large gap in our understanding of trust. Using a basic unit of trust (*S trusts T to ϕ*) as a framework, I focused in this paper only on *kinds* of trust and describe three, based on the way we develop a relation of trust with another agent: communication-based trust, practice-based trust, and context-based trust. Given these three kinds, I offered an answer to the question: what kinds of agents can become trustees for each kind? In this way, I attempted to align kinds of receivers of trust with kinds of trust, arguing that not every kind of potential trustee can be trusted in every kind of way. Because human trust is such a deep and rich topic for philosophic analysis, there is much more work to do on these and on other possible kinds of trust; but I believe that I have provide enough of a start on several important kinds of trust to be able to move to part II of my project: the degrees and forms of trust.[20]

Notes:

1. For this section of my paper, I rely in large part on the analysis of trust in Carr (2012). I note that this is practical, as opposed to epistemic, trust; it is trust between practical agents with respect to an action and the practical value it has for the trustor, it is not trust between epistemic agents with respect to information and the truth-value it has

for the trustor. In addition to Carr (2012), see, e.g. Macleod (2011), Hardin (2002, p. 9), and Hardin (2006, p. 19) for descriptions of the standard model.

2. It is not only by stipulation that trust is here taken to be a complex, primarily cognitive, mental state (of a human agent) composed of three interrelated mental acts; I believe that this is an intuitively plausible conception of trust that will gain strength as differences between the forms, kinds, and degrees of trust are explored. It is standard in the trust-literature to connect trust to an expectation and to belief/knowledge concerning the trusted's trustworthiness; concerning expectation, Hardin (2006, p. 29) remarks, "In virtually all conceptions of trust, there is an element of expectation." With respect to the decisional or intentional element to trust, I note that in her Tanner Lectures Baier (1991) makes a point of stating that trust is rarely a decision (p. 123); she focuses on trust more as a response, initiated by the trusted who requests to be trusted or indicates he can/should be trusted. This claim (to me an overstatement) is in line with Baier's attempt to counter rational choice (game-theoretic) theories of trust; but if trust is not to be an *involuntary* response, it is hard to see how an intention or decision on the part of the trustor is still not an important part of establishing a trust relation. Likewise, Hardin (2006, chapter 2) argues that trust is not a choice or decision; he means by this, however, that an agent cannot simply "will" or "decide" or "choose" to trust another agent in the absence of epistemic input concerning that agent (which I cover by the third mental act: belief).

3. See Macleod (2006) for reference to the literature on trust offering a variety of such definitions of trust. As I see it, the problem with defining trust in terms of a disposition, a character trait, or an abiding contribution to one's personality, is that these are ordinarily thought of as psychologically invariable relative to an agent's changing situations; it makes trust a property of the trustor (S) who becomes a "trusting" type of person (either dispositionally, or in character, or in personality) whether the situation is one in which this agent actually trusts or not. And it implies that distrust is a property of the dis-trustor who becomes the psychologically opposite "type" of person, again no matter the situation. This, it seems to me, is the wrong metaphysics; trust seems less a property of an entity (person) and more a relation between entities (persons) - even though logic treats an n-ary relation as an n-place predicate (property). In addition, methodologically such definitions tend to shift focus away from trust as a relation between agents to trust as a property of a person. In contrast, the definition I will be using does not characterize a personality "type," but makes the act of trusting a relation from one agent to another that is strictly relative to a particular situation: a given trust-complex; external to the given situation, the agent is no longer *that* trustor.

4. Of course, if T is a human agent who knows she is being trusted by S to ϕ, and on this basis informally promises S or enters a formal contract with S to ϕ, and if ϕ'ing is a socially, legally, morally permissible action, then T takes on the obligated to ϕ; but T's duty in this case derives from the promise (or contract) and the permissibility of ϕ'ing, not from the mere fact that, in trusting T to ϕ, S forms an expectation that T will ϕ. Obligation would follow only where: (a) the trusted is an agent with sufficient rationality (e.g. the average adult human), (b) the action is permissible, (c) the trusted knows and accepts the trust she is given, and agrees to do the action in question.

5. That trust is normative or contains elements of normativity has been pointed out and argued by several authors; see Baier (1986, 1991), Macleod (2011), and Buechner and Tavani (2011).

6. That trust can be a morally good or bad thing, requiring its own moral evaluation, is widely held in the trust literature. See, for example, Baier (1991) who, in her Tanner Lectures "Trust" is largely concerned with such moral evaluation. See also Gambetta (1988, 213ff) for a forceful statement on the moral neutrality of trust within cooperative interactions.

7. Williams (1988), for example, argues that dependency is a central element in trust.

8. As noted by Macleod (2011), the trust literature acknowledges and confirms our ordinary experience that there is an element of risk in trust. There is, however, an influential thesis in the trust literature, held by Luhmann (1988) and Gambatta (1988) among others, that there is a trust relation only in cases where what the trustor risks is *greater* than what the trustor values. So, if S trusts T to ϕ, the value S gives to ϕ'ing must be *less* than the value to S of what S foregoes should T fail to ϕ; otherwise S does not *trust* T to ϕ, S simply makes a rational choice whether or not to rely on T to ϕ. To illustrate this thesis: suppose you are considering eye surgery to be done by a surgical robot. The result of a successful operation is moderately improved vision, the result of a failed operation is permanent blindness, and no operation means that you continue to have slowly declining poor vision. According to the above thesis, if you value moderately improved vision more than the value you give to not being permanently blind, then you calculate the expected utility of each outcome (moderately improved vision, permanent blindness, and slowly declining poor vision) and make a rational choice to have the operation or not. If the rational choice is to undergo the surgery, you do not *trust* the surgical robot; you (presumably) "rely" on it to perform the operation successfully. But if you value not being permanently blind more than you value moderately improved vision, and you still decide to undergo the operation, then your relation to the surgical robot is one of trust to perform a successful operation. I disagree with this thesis for interpersonal trust, but it seems to have merit worth exploring for interagential trust. However, the brief formulation of the connection between trust and risk to which this note is attached is not intended to take a stand one way or the other.

9. As noted earlier in this paper, this last part of the trust relation (the trustor's belief that the trustee is relevantly trustworthy) will be examined in part II of my project to distinguish kinds, degrees and forms of trust.

10. Annette Baier (1986), for example, in her influential work on trust focuses primarily on single human trustees; she includes trust of and within institutions and levels of administrations in her Tanner Lectures (1991). Russell Hardin (2002) likewise takes "dyadic interpersonal trust" as basic and applies insights from this case to trust of government (Ch. 7) and trust of society Ch. 8). Trudy Govier (1998, pp. 14ff) follows the same method; after presenting 2-person interpersonal trust as basic trust, she proceeds to extend it to a variety of potential trustees, including inanimate objects and dead humans. deVries (2010) clearly expresses and works within the same intuition.

11. Govier (1998, pp. 14ff) extends trust to such technologically organized systems of human activity. For a theoretically detailed and explicitly argued position that extends trust to artificial agents that function within human "contexts" or "zones," see Buechner and Tavani (2011).

12. Compare Govier (1998, p. 8), who connects trust with communication, but differently than I present it here. My interest is trust that comes about by communication, while Govier discusses communication that depends on and takes place within an already formed trust relation.

13. Simon (2013); see Section "Trust and Trustworthiness."

14. I use "practice" in the sense it has in political and social philosophy, not the sense it has in education. A "practice" is a pattern of behavior or interaction, publically recognized, governed by formal (e.g. legal regulations) or informal (e.g. cultural codes of behavior) social norms, that functions as (or has the status of) a practical convention (i.e. it has its stability and is maintained by people continuing to engaging in it). A person's job or profession, patterns of exchange (face-to-face or on-line) in our economic lives, routines of social behavior in which people engage, and activities related to citizenship such as voting or partaking in civic celebrations are easily recognized examples of practices.

15. This section draws generously from the work of Buechner and Tavani (2011) and especially Tavani (2014 forthcoming).

16. This characterization of "form of life" comes from Saul Kripke's (1982, pp. 96ff) understanding of Wittgenstein's notion of a "form of life."

17. Compare Govier (1998, pp. 14ff) for trust extending into the past, and trusting the dead. It is, on the one hand, surprising that (present) trust of future HAs is not a major topic in the trust literature; it would seem that many important long-term human projects would not happen without trusting future generations to see them through, and that future generations would not see them through without trusting the HAs (to them long dead) who initiated them. On the other hand, however, it is perhaps understandable that a 2-term relation (e.g. trust) of which one term does not (yet) exist might be thought not a possible topic of study.

18. Care must be taken to understand the subtle argument at the heart of the Buechner-Tavani contextual theory of trust correctly. It would be the fallacy of division if they are in effect arguing: (1) Each zone of DDT is a whole. (2) DDT extends to the whole zone. (3) AAs are part of this whole. (4) Therefore, these AAs are trusted. Also, given that trust is non-transitive, they can't be arguing: (1) Within a given zone of DDT, HA_1 trusts HA_2. (2) HA_2 trusts other HAs and AAs within this zone, and they in turn trust … . (3) Therefore, HA_1's trust propagates to every agent within that zone. Thus, in their theory DDT is "diffuse" not by whole-part reasoning and not by transitivity reasoning; it is "diffuse," it would appear, by the quality of the trustor's life as the trustor occupies a given zone with "ease," "comfort," and "safety."

19. It is one thing for a motion sensor to detect the *absence* of motion, quite another for it *not* to detect motion; it must be designed to do the first, but only designed to detect motion to "do" nothing when not detecting motion. A light that is off because it is not designed to detect no motion (i.e., designed to be on when detecting motion) is not the same as a light designed to be off when it detects no motion; off is its default state in the former, but not in the later.

20. This paper has benefited from several critical comments and suggestions from my colleague Herman Tavani, for which I am grateful.

References:

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96, no. 2: 231-260.

Baier, Annette. 1991. "Trust." Tanner Lectures on Human Values. Princeton University, March 6-8, 1991. Available at: **tannerlectures**.utah.edu/_documents/a-to-z/b/**baier**92.pdf

Buechner, Jeff and Herman 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.

Buechner, Jeff, Judith Simon, and Herman T. Tavani. 2013. "Re-Thinking Trust and Trustworthiness in Digital Environments." In. G. Costa, ed. *Ambiguous Technologies: Proceedings of the Tenth International Conference on Computer Ethics – Philosophical Enquiry*. (July 1-3, 2013). Autónoma University, Portugal.

Carr, Lloyd. 2012. "Trust – an analysis of some aspects" pdf. Available under *Writings* at: http://rivier.edu/faculty/lcarr/

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

Gambetta, Diego. 1988. "Can We Trust Trust?" In Gambetta, Diego, ed. 1988. *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell.

Govier, Trudy. 1998. *Dilemmas of Trust*. McGill-Queens University Press. Montreal, QC, CAN.

Hardin, Russell. 1990. "Trusting Persons, Trusting Institutions." In Zeckhauser, Richard, ed. 1991. *Strategy and Choice.* MIT Press. Cambridge, MA.

Hardin, Russell. 1996. "Trustworthiness." *Ethics*: 107, 1 (Oct. 1996) 26-42.

Hardin, Russell. 2006. *Trust: Key Concepts*. Polity Press, MA

Hardin, Russell. 2004. "Distrust: Manifestations and Management." In Hardin, Russell, ed. 2004. *Distrust. Vol. VIII, Russell Sage Foundation Series on Trust.* Russell Sage. New York.

Hardin, Russell. 2002. *Trust and Trustworthiness. Vol. IV, Russell Sage Foundation Series on Trust.* Russell Sage. New York.

Kripke, Saul. 1982. *Wittgenstein On Rules and Private Language.* Harvard University Press.

Luhmann, Niklas. 1988. "Familiarity, Confidence, Trust: Problems and Alternatives." In Gambetta, Diego, ed. 1988. *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell.

March, Stephen Paul. 1994. *Formalizing Trust as a Computational Concept*. PhD Dissertation, University of Stirling. Available at: www.no.nr./~abie/Papers/TR133.pdf

McKnight, D. Harrison and Chervany, Norman. 1996. "The Meanings of Trust." Available at: misrc.umn.edu/workingpapers/fullpapers/1996/9604_040100.pdf

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

Simon, Judith. 2013. "Trust." In Oxford Bibliographies. Oxford University Press. Available at: www.oxfordbibliographies.com

Tavani, Herman T. and Jeff Buechner. 2013. "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.

Tavani, Herman T. 2014 (forthcoming). "Degrees of Trust in the Context of Machine Ethics."

Williams, Bernard. 1988. "Formal Structures and Social Reality." In Gambetta, Diego, ed. 1988. *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell.

(January, 2014)