Axiomatic and Behavioural Trust

Clark Thomborson

Department of Computer Science
The University of Auckland, New Zealand
cthombor@cs.auckland.ac.nz

Abstract. Academic discourse on trust is fractured along disciplinary lines. Security theorists routinely use a definition of trust which, apparently, has little in common with any of the definitions of trust that appear in the sociological and psychological literature. In this essay, we extend a recently-proposed framework for the technical analysis of secure systems, so that its notion of trust is roughly congruent with the sociological theories of Parsons, Luhmann, Barber, Lewis and Weigert. This congruent extension suggests some ways in which a computerised system might, appropriately, inspire trust in its non-technical users.

Key words: Trust, security analysis, trust management

1 Introduction: Two Types of Trust

My recently-proposed security framework [1] provides terminological and definitional support for security analysts in diverse subfields, retaining the common elements of their modelling approaches in the framework while excluding the subfield-specific detail. The definitions and taxonomic classifications in the framework are, for the most part, cross-products of dichotomised variables. For example, the four concepts of security, functionality, trust, and distrust are quadrants in the two-dimensional space defined by two binary variables: feedback and assessment. Functionality and trust involve positive feedback to the owner of a system, whereas security and distrust involve negative feedback. Trust and distrust, as defined by Luhmann [2] and others, are not based on assessment – instead they are characterised by an uncertainty or lack of information about some (presumed, relied-upon) "good" or "bad" contingency. Functionality and security, by contrast, are an owner's assessments of likely future positive or negative feedbacks from their system.

Our approach to the understanding of trust can be viewed as a harshly simplified version of a functional, cybernetic sociological theory. As such, it cannot offer any startling new insights to sociologists, but our reductions and simplifications may be helpful in clarifying the definitions and distinctions made in the conceptual models developed by Luhmann, Barber, and others. However the main goal of this essay is not to contribute to the sociology of trust, but instead to offer support for interdisciplinary discussions of the nature and functions of trust. Technologists can gain useful insights about trust, we argue, from

the prominent sociologists of the latter part of the twentieth century. We also suggest, from our technological perspective, some experimentation that sociologists and psychologists might conduct, if they wish to elucidate the structural foundations of trust and the primary factors in an individual's trusting decisions.

We start our exposition by distinguishing two subtypes of trust within our framework. This distinction will allow us to discuss the most important mode of secure-systems analysis, wherein some range of system behaviour is proved to be impossible. Such security proofs are rigorous deductions on a set of axioms, where each axiom constrains the behaviour of the "trusted" elements of the system. Theorems are of the following form: no matter how the untrusted elements of a system (mis)behave, the system will still have some desirable security property. In the context of such proofs, "trust" is thus an axiomatic property, one which is ascribed to certain elements of a system by its modeller. In order to reconcile this notion of trust with our framework, we must define two subtypes of trust.

Behavioural trust statements are either made by, or ascribed to, a set of actors in a model. Such statements are confident (but still somewhat uncertain) descriptions or predictions, by a *trust analyst*, of some desirable behaviour by one or more actors in the model.

Axiomatic trust statements are axioms of a system model, defining the desired behaviour (as judged by the owner of the system) of one or more actors in the system. Axioms are uttered, or assented to, by a *trust modeller* who is, at least notionally, external to the model defined by the axioms.

We have identified four ways in which an actor in a modelled system may gain the externality required to act as trust modellers within their own system. The modeller who created their system may have provided an *oracle* which provides axiomatic advice to actors in a model, in some language that can be interpreted by these actors. Any actor in a system may construct a *subsystem* along axiomatic lines. Any actor in a system may adopt an *induced axiom* by uttering or assenting to statement which they hold to be true beyond reasonable doubt. If an apparent contradiction arises, the contrary evidence is investigated carefully before the axiom itself is questioned. An example of an inductive axiom is $E = mc^2$. Finally, an actor may derive some *deduced axioms*, i.e. lemmas or theorems, from their current set of axioms by a process of logical deduction.

We note that axiomatic trust statements are formally correct. Their validity is questioned only when a set of axioms is discovered to be logically inconsistent by a novel deduction, or when a novel set of observations (e.g. of the behaviour of photons near the Sun) clearly invalidates an induced axiom such as Newton's law of gravitation. By contrast, the validity of every behavioural trust statement is formally uncertain – it is directly contingent on future observations. This is Luhmann's distinction between "confidence" and "trust" [3]. Luhmann also defines the meta-axiomatic concept of "familiarity", in order to discuss the language in which the modeller expresses her axioms. The modeller has "confidence" in the validity of their axioms; anyone who formally analyses this modeller must base their analysis on axioms which describe the modeller's "familiar" language. We do not adopt Luhmann's triad of definitions because they are more detailed than

we require for our discussion, and because we have already defined trust in a broader way: as the system owner's unassessed expectation of positive feedback from their system.

We distinguish distrust from trust, by considering the difference between the fears and desires of the owner of the system. Although this essay is devoted to an exploration of trust, we define, in passing, an *axiomatic distrust statement* as an anti-requirement on a system, specifying what it shall not do. A *behavioural distrust statement* is the expectation of an system analyst, regarding the likely "bad" behaviour of one or more actors in that system. Clearly: an analyst who distrusts some behavioural aspect of an actor will endeavour to avoid depending on that actor in the relevant contexts. Such active avoidance, when enacted, becomes a security provision, that is, a system modification whose expense or inconvenience is justified by the owner's assessment of a reduction in harm. As such, a distrusting decision is clearly distinguishable from the functional motivation of a trusting decision, and is deserving of a separate analysis. Later in this essay, we will return to the issue of decision-making with respect to security (costs or other harms), functionality (benefits), trust (uncertainty about benefits), and distrusts (uncertainty about harms).

We define a *model* to be any simplification of a real-world system which is too complex for a direct analysis. A competent modeller will search for radical simplifications which make their model maximally analysable, while doing as little damage as possible to the accuracy and precision of the model predictions. A logically-inclined modeller defines a model by constructing its axioms. An experimentalist defines a model by constructing it from the material at hand, that is, from the malleable elements of the system in which the experimentalist is an actor. We imagine that almost all of the models people use in their everyday lives are of the experimental variety: they are incompletely axiomatised. However every model, as defined here, has a purpose of simplified representation; so we argue that every modeller has uttered at least one (perhaps only vaguely apprehended or understandable) axiom when constructing their model.

Luhmann asserts that everyone in our ever-more-complex modern world is, increasingly, engaging in such a process of model-making and analysis, because such analytic simplifications are required to thrive and perhaps even to survive:

> We are now in a position to formulate the problem of trust as a gamble, a risky investment. The world is being dissipated into an uncontrollable complexity; so much so that at any given time people are able to choose freely between very different actions. Nevertheless, I have to act here and now. There is only a brief moment of time in which it is possible for me to see what others do, and consciously adapt myself to it. In just that moment only a little complexity can be envisaged and processed, thus only a little gain in rationality is possible. Additional chances of a more complex rationality would arise if I were to place my trust in a given future course of action of others (or for that matter in a contemporary or past course of action, if I can only establish it in the future). If I can trust in sharing the proceeds, I can allow myself forms of co-operation

> which do not pay off immediately and which are not directly visible as beneficial. If I depend on the fact that others are acting, or are failing to act, in harmony with me, I can pursue my own interests more rationally—driving more smoothly in traffic, for example [2].

We do not attempt to express all of Luhmann's theory of trust in our framework, however we have adopted what we see as the primary elements of his theory. In particular, we do not insist that the primary motivation for every behavioural trust statement is the Luhmannian purpose of uncertainty reduction during decision-making. Such an insistence would limit the scope of our modelling to actors who are sentient and purposeful entities. We see no analytic advantage – and we see some disadvantages – in ascribing a purpose to a computerised actor which is making predictions about future events. We survey a few other theories below, regarding the purpose or function of trust.

In Parsons' AGIL theory [4], every group has a primary functional imperative of pattern maintenance. If the axioms of Parson's theory are adopted (as axiomatic trust statements) in a model within our framework, then a social group's self-descriptions are behavioural trust statements about itself. These self-descriptions are an emergent behaviour of the group, and are developed by intersubjective processes. Every group is expected, by a Parsonian analyst (in our radically simplified model of Parsonian theory!), to mount a spirited defense if the validity of a self-descriptive statement is questioned or threatened.

In a game-theoretic analysis, a behavioural trust statement is a confident description of an player's current tactics and strategy. The rules of the game itself are axiomatic trust statements. If the analyst is acting as a player in the game, then their behavioural trust statements have a clear Luhmannian purpose of uncertainty-reduction. However if the analyst has some other motivation for their analysis, for example intellectual curiosity or a desire to help a player improve their game-playing abilities, then the Luhmannian purpose of uncertainty-reduction seems an insufficient motivation for the analysis.

In some computer-mediated economic markets such as eBay, behavioural statements are automatically generated to express the market-controller's radically simplified estimate of the "reputation" of a vendor or buyer. Participants in such markets are encouraged to rely on such statements, when deciding whether or not to take the risk of closing a transaction. The primary purpose for eBay's utterance might be profit-expansion rather than uncertainty-reduction, but until a teleological system is specified, such speculation is pointless.

A modelling framework must be agnostic on disputed axioms, if it is to aid communication between modellers who have different belief structures or goals. For this reason, we do not presume that a axiomatic trust statement must be consciously constructed by an analyst before it is used by an actor in some model. Instead, the analyst may construct the axioms describing a model of an actor (or a set of actors) on the basis of their prior behaviours.

Similarly, we do not insist that an analysis must be conducted along rational lines. Depending on the axioms in the model, the analysis might be conducted along emotive or faith-based lines. Furthermore, the analysis might, or might

not, be conducted by a single person. A psychological analysis of trust requires a model whose axioms support an analysis of individual actor's trusts. A valid sociological analysis of trust, even if is axiomatised, requires an intersubjective process to develop and interpret the axioms.

We now review the theoretical frameworks of a few prominent sociologists, not in the hope of finding an undisputed set of axioms for a system analysis, but in order to discover whether our framework can help elucidate and harmonise these theories, and whether we can discover any inconsistencies or limitations.

2 Barber's Subtypes of Trust

Barber, in his influential monograph, identified three fundamental forms of trust statements [5]. His "general and comprehensive definition of trust" includes an "expectation of the persistence of the moral social order." He also offers "two more specific meanings [of trust], each of which is important for the understanding of social relationships and social systems": an expectation of technically competent role performance, and an expectation of fiduciary obligation and responsibility.

We note that Barber's general definition is an axiomatic trust, under the presumption that a general collapse of the moral and social order is just about unthinkable. His specific meanings are behavioural trusts, because they are reliances on specific individuals or organisations. It seems possible for an individual to place axiomatic trust in an institution such as the Catholic Church, even though they distrust the moral and social order of their immediate environment. Thus Barber's theory of trust seems questionable as a psychology of trust, but his general definition seems an appropriate axiomatisation of a societal trust, and his specific definitions are a subtyping of our behavioural trust.

We now briefly describe the structural (morphological) aspects of our framework, for this is necessary before we can describe how our framework supports a Barberian analysis of a trustworthy profession.

In our framework, a system model consists of a set of actors A in a network of pairwise relationships R between actors. That is, every model is a graph with vertices A and edges R. If an analyst wants to employ our framework to perform a Barberian analysis of the trusting behaviours of a particular political system, a profession, or a family, they should start their modelling procedure by identifying a representative set of actors and relationships.

Only three types of relationships are defined in our framework. An actor is *superior* to another actor, when the analyst wants to express the power that the first actor has to observe and control the second actor. An actor is a *peer* of another actor, when the analyst wants to model the friendly conversations and consensual agreements which can be made between actors who – at least in this context – have no great imbalance of power. Finally, an actor is an *alias* of another actor, when the analyst wants to model either the multiple role-playing abilities of a person, or the multiple aspects of a subsystem (whether it be mechanised, in whole or in part) depending on its observer and the context.

An adequate model of a complex social arrangement, of the sort diagrammed by Granovetter, has multiple aliases for each person who is represented in the model: one alias for each of their interpersonal relationships.

We consider a specific example to illustrate the modelling process. An idealised professional society can be modelled as a set of peers who, collectively, enforce some membership criteria. Any peer who has been found to violate these criteria can be expelled, or they may be chastised (perhaps only in private conversation within the peerage) and allowed to retain their membership. It is in the collective self-interest of the professionals to be considered, by the general public, to be trustworthy. The professionals may regulate themselves effectively; they may have a laissez-faire attitude; and they may be a solidary group which actively defends any member against an ethical complaint raised by one of their clients. Since a client is not able to monitor the conversations among the professional peerage, they have no way of knowing, for certain, whether the peerage is effectively enforcing any fiduciary responsibilities or technical competencies. However if a potential client trusts the professional peerage to enforce these two Barberian "specific meanings" of trust, then (if they follow the Barberian axioms when making their trusting decision) this client will confidently accept professional services from any member of the peerage.

Formally, a profession is a system with at least two sets of actors: the professionals P and their potential clients C. Barber's analysis of trust in the professions includes a third type of actor, a government with jurisdiction over the professionals and clients. When modelled in our framework, a government is an actor that is superior to each professional and client. Because the professionals are not acting as functionaries in the government, we introduce aliases for the professionals and clients so that we can represent the type of power (possibly only an informational advantage) that is wielded by a professional over their client.

Our complete model is shown in Figure 1. The professionals (p, q) and client c, each have a primary alias (p_0, q_0, c_0) that has power over itself – these arcs represent the self-control and self-reflection abilities of any sentient individual. We indicate that c is accepting professional services from p by introducing aliases c_p and p_c, with p_c being a superior to c_p. Conflicts of interest may arise in any aliased relationship. For example, p_c may give advice to c_p which c_0 is not happy to accept; and c_p's relationship with p_c may put p_c (the professional persona of p) into conflict with p_0's self-interest.

The professional society is modeled as a peerage (on aliases p_p and q_p) in our framework. The line connecting the peers is intended to suggest a communication network, such that anyone on the network can send messages to anyone else on the network. Peerages, if they organise themselves, can develop a way to form a collective opinion, possibly using some subsidiary actor or device (drawn below the network) to collect the votes. In the case of Figure 1, the peers have collectively hired a Lobbyist, who is under the control of the peers and transmits messages on their behalf to an alias G' of the government. The Lobbyist in this Figure is apparently insentient, for it has no self-control arc – it is may be just

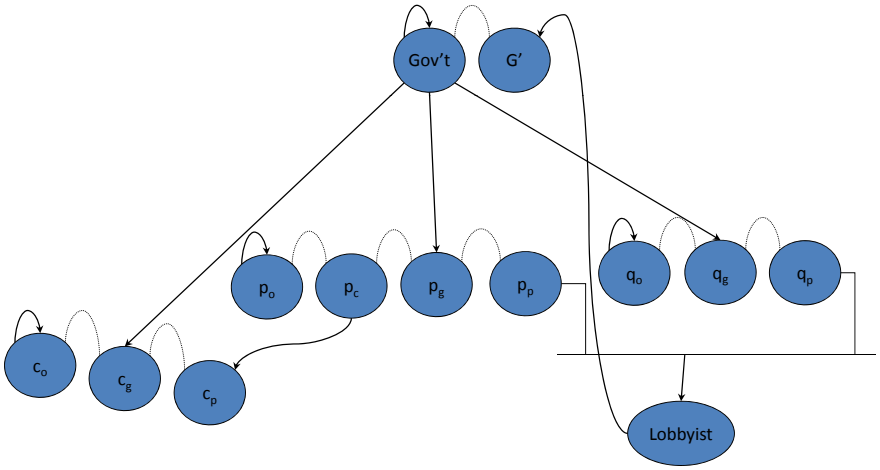

Fig. 1. A client, two professionals, a lobbyist, and a government. The professional peerage consists of actors p_p, q_p, and the Lobbyist. The governmental system is a hierarchy, with c_g, p_g, and q_g as inferior actors. Some amount of self-control, i.e. some form of free will, is possessed by the personal actors c_0, p_0, q_0, and the group actor Gov't. The Gov't may accept advice from the Lobbyist, and the client c_p may accept advice from the professional p_c.

an email-forwarding system. Alternatively, the modeller may be simplifying an already complex system by suppressing the alias representing the self-interest of the Lobbyist, because it is not relevant to the modelled situation.

In our framework, peers have no direct power over each other, aside from ostracism from their peerage. In the system of Figure 1, peers have some indirect power over each other, because they can lodge a complaint against a peer to the Government through their citizen alias p_g or q_g. Also, peers might petition the government for an intervention via their Lobbyist – if the other peers do not prevent the transmission of this petition via their (partial, shared) control over the Lobbyist.

We are now ready to discuss Barber's analysis. He points out that a client's trust in the fiduciary or technical behaviour of the profession P is increasingly problematic, for several reasons. For example, the Lobbyist for the professional society may exert inappropriate control over the government G, through some branch G' of the government that is effectively under the control of the Lobbyist. The governmental branch G' in Figure 1 is thus a distrusted actor, from the perspective of the client c; but it might be highly trusted by the peers.

Barber argues that "trust alone is not enough to ensure [the professionals'] effective performance in the public welfare." In terms of Figure 1, Barber is arguing that the client must, somehow, be assured that the Lobbyist is not effectively in control of the government, and that their government is competently discharging a fiduciary responsibility by regulating the peerage. Barber indicates,

by example, some ways in which these assurances can be made. It is impossible to do justice to his essay in this brief summary. However, roughly speaking, his general prescription is for professional groups to accept governmental regulation as a counterbalance to professional self-interest. In a revision to Figure 1, we might follow Barber's advice by adding a Regulator that is inferior to the Gov't. This Regulator, to be effective, must have at least some observational power over the peerage. If a peer misbehaves, then the government can punish the guilty citizen (p_g or q_g). The Regulator may, in addition, have some direct control over the decisions of the peerage. The Regulator can be given full observation rights by granting them a non-voting membership in the peerage; and the Regulator can be given veto rights over the Lobbyist by making the Lobbyist an inferior of the Regulator rather than an inferior of the peerage (as in Figure 1). However if the Regulator has any observation or control power, the professionals must trust the Regulator not to abuse this power, for example by inappropriately vetoing a proposed action of the peerage, or by inappropriately revealing a peer's confidential information to Government (and potentially to the general public through the Government's control over the citizenry). Barber argues that professions will enjoy an increased level of trust from their clients by a well-crafted governmental regulation, if it is openly and fairly administered. Clients may be expected to seek (and to follow) professional advice much more readily if its source is trusted.

We conclude that our framework is adequate to capture the main line of Barber's discussion about trust in the professions, and that our structural diagrams may be used to clarify the ways in which a governmental Regulator could monitor or control a profession.

Our framework provides only marginal support for Barber's three types of trust. His general and comprehensive definition of trust is roughly congruent to our axiomatic trust, but seems problematic (as a "general and comprehensive" definition) in any case where an individual places little trust in the moral and social order, but does trust specific individuals or institutions. Barber's "two more specific meanings" roughly fit within our definition of behavioural trust, with expected behaviours of technical competency being in one of Barber's subcategories, and expected behaviours of fiduciary obligation and responsibility being in the other.

Inferiors must trust superiors not to abuse their power. Barber's subtyping suggests that one form of abuse can be ascribed to an inappropriate control of a superior's other aliases (especially their self-interest persona, e.g. p_0 of the professional p in Figure 1) over the actions taken by their superior-role persona (e.g. p_c). The other type of abuse (insufficient technical competency) can be ascribed to an inadequately provisioned superior-role persona.

Peers must trust each other. Barber's typology will help us distinguish conflicts of interest (where a peer's duty to a peerage is contrary to their self-interest or other responsibilities, i.e. to their government) from the functional inadequacy of a peer.

Individuals must trust themselves. Barber's typology is a sociological theory rather than a psychological one, so some difficulties may be expected in this extension of his theory. Even so: Barber's distinction suggests the following psychological analysis. An individual's trust in their personal competency (e.g. as determined by p_0 in the case of our professional p of Figure 1) can be distinguished from their trust in their management of their conflicts of interest (as determined by all the other aliases of p in Figure 1). When the self-trust in competency is violated, an individual is unable to help themselves (or to refrain from hurting themselves) due to some personal inadequacy. When the conflict-of-interest self-trust is violated, an individual is in the painful position of being expected (by external controllers and observers) to harm themselves (or neglecting to help themselves) in order to discharge their responsibilities to others. This is the fundamental conflict of humanist psychology, but as far as we know, it has not been previously identified as arising in Barber's theory of trust.

We conclude that Barber's subtyping distinctions can be made in our framework, if the axiomatic system of the model clearly distinguishes an individual's responsibility to themselves from their responsibilities to others, and if personal inadequacy can be distinguished from externally-imposed constraints. The latter distinction is commonly made but problematic, in the highly reduced context of human-computer interaction. An unpleasant incident must be attributed to either operator error or to a poorly-designed (or buggy) program, before operator training or program revision can be chosen as the more appropriate resolution.

3 Emotional and Rational Trust

Lewis and Weigert argue that trust has "distinct cognitive, emotional, and behavioral dimensions which are merged into a unitary social experience" [6]. In this essay, we have argued that the behavioural dimension of trust is a primary consideration in a framework, such as ours, which prescribes a structural approach to modelling. However we note that the antithesis of behavioural trust is the axiomatic trust required, in the observer of another actor's behaviour, to ascribe any meaning to, or consequential implication of, this behaviour. Our axiomatic trust is essentially an attribute of the observer, and our behavioural trust is an attribute of the observed. The dichotomising variable is the viewpoint: observer versus observed.

The cognitive and emotional dimensions of trust, as defined by Lewis and Weigert, are complementary. They demonstrate this complemantarity by imposing a three-level scale on each of these two dimensions, forming the 9-category cross-product of Table 1.

We would recommend this table to any technologist who is trying to persuade a non-technical person to use a system. In our experience, technologists tend to operate in the top row of this table, although some will bristle if they are "accused" of having an ideological trust in their analytic methodology or results. By contrast, anyone who is unable to understand a technical analysis is limited to the bottom row of the table, whenever they consider using a technically-complex

Rationality \ Emotionality	High	Low	Virtually Absent
High	Ideological Trust	Cognitive Trust	Rational Prediction
Low	Emotional Trust	Mundane, Routine Trust	Probable Anticipation
Virtually Absent	Faith	Fate	Uncertainty, Panic

Table 1. Nine types of trust, classified by emotionality and rationality [6].

system. A technologist is likely, in my experience, to adopt a strategy of cognitive argument: essentially attempting to educate the potential user so that they can operate in the second row in Table 1. Education is certainly appropriate, if the system is not sufficiently foolproof to be used safely by a complete novice. However, if the technologist operates in a requirements-elicitation frame, rather than in a strictly educative frame, then the potential user's desires and fears may reveal some novel requirements for their system and especially for its user interface.

The sociology of distrust has received much less attention than the sociology of trust, even though many theorists have argued that it is best treated as a separate topic. We note that Lewis and Weigert have characterised the near-absence of both cognitive and emotional trust as "uncertainty, panic". We hypothesise that the level of distrust is what distinguishes a state of uncertainty from a state of panic. If the prospect of continued inaction is distrusted intensely, and none of our options for purposeful action are trusted, then energetic but unpurposeful behaviour is the best option under a cognitive analysis, and a state of panic seems the most likely emotional response. However, in a situation where inaction is not distrusted, a lack of trusted options for the next action is not a cognitive stress. The emotional status of our hypothetical individual seems underdetermined, but one possibility is that they are taking a serene step on their path to Nirvana. We conclude that Lewis and Wiegert's intriguing table offers fertile ground for future experimentation regarding the interactions of trust and distrust, in their cognitive and emotional manifestations.

4 Trust, Distrust, and Decision-Making

In this section, we state and explore a hypothesis about the psychology of a human decision. This hypothesis is grounded in the taxonomic theory of our framework, as extended here. We invite correspondence from sociologists, psychologists, and market researchers who can point us at any article in which some variant of our hypothesis has been validated or invalidated.

Within our framework, it is natural to model rational decision-making as occurring on three dimensions. On the economic dimension, the decision-maker

will assess whether the expected benefit of an option exceeds its expected cost. On the optimistic dimension, the decision-maker will assess whether their confidence (a behavioural trust statement) in the favourable outcome of an option is sufficient to overcome their fundamental uncertainty about this favourability. Finally, on the pessimistic dimension, the decision-maker will assess whether their sense of control over their expected future status exceeds their level of distrust about this expected status, for each of their options.

We hypothesise that options are considered serially, with a binary go/no-go decision taken on each. There may be multiple rounds of decision-making before a visible action is taken, depending on the perceived urgency and importance of the decision. Experimentally, it will be simplest to work with single-round (i.e. urgent or unimportant) decision-making before attempting to unravel the complexities of a multiple-round decision.

Any option with an expected net benefit, which is sufficiently trusted, and whose analysable downsides are sufficiently controlled, would (we presume) provoke a "go" decision. By contrast, if an action has a poor benefit, is insufficiently trusted, and has many uncontrolled downsides, then it seems a clear "no-go".

If there is any disagreement on the three dimensions, then two of the three assessments must agree. We postulate that a decision that is unfavourable on two dimensions would never be taken, except when the decision-maker is in a state of panic. We would define this state to arise when the prospect of inaction is highly distrusted. A panicked individual will, we presume, take the first option for action unless it is even more distrusted than the status quo.

If there is only one argument against taking an action, and the status quo is not distrusted, we would expect some humans to take the action. The outcome would depend, we presume, on a personality type. Economists will weight the first dimension most heavily; Optimists will pay most attention to their attractions; and Pessimists will pay most attention to their fears. We doubt that this personality typology is novel, since it seems rather obvious; but we are not aware of the relevant experimental literature.

To illustrate a possible use of our hypothesis, we imagine that we are advising a technologist who has designed a system which they firmly believe would be beneficial for most people to use. The technologist seeks our help in understanding why so few people are using their wonderful system. We would advise this technologist to offer twenty-five randomly-selected people a chance to use the technology, and to classify their responses into five categories. Under our decision-making hypothesis, we would expect the classification to be unambiguous except in a few cases where the prospective user expresses multiple reasons for deciding against using the system.

Category -1: The prospective user refuses to use the system, and the technologist is unable to classify their reason for refusing.

Category 0: The prospective user decides to use the system.

Category 1: The prospective user decides against using the system, because they see no expected net benefit in using the system.

Category 2: The prospective user decides against using the system, because they have insufficient trust in the system.

Category 3: The prospective user decides against using the system, because they distrust it.

If category -1 is frequent, we would advise the technologist to engage a more emotionally-communicative interviewer. We would also question the validity of our theory of decision-making, as well as the efficacy of our instructions to the technologist.

If category 0 is frequent, we would advise the technologist to review their advertising campaign. The current advertisements may be insufficiently distributed; they may be much less effective than the direct-sales approach (as conducted by the technologist); and they may be targeted at an audience that differs significantly from the population the technologist interviewed.

If category 1 is frequent, the technologist should attempt to improve the economic performance of the system, and they should look for ways to communicate this performance to their prospective but non-expert users.

If category 2 is frequent, the technologist should attempt to develop a radical simplification in the user interface of our system, so that the prospective user is not faced with a difficult task when developing their own behavioural trust statements about the system. As noted by Luhmann and many others, statements of the form "trust me" or "trust my system" unlikely to increase trust, for they tend to call trust into question. Trust is fundamentally non-assessed. Descriptions of how well the system will respond under adverse conditions (malfunctions or attacks) are arguments against distrust, not arguments for trust.

If category 3 is frequent, the user's distrust is the key variable. If the fears are articulated sufficiently clearly that the relevant system behaviours can be understood by the technologist, then the technologist can look for ways to demonstrate that the system is foolproof in these respects. If the system is not already foolproof against such malfunctions, then improving the design in this respect should be a high priority. If the technologist's search for a clearly-demonstrable architectural control proves infeasible, then economic, social, or legal safeguards should be explored. It will be important to know, at this juncture, whether our prospective users are, generally, trustful of architectures, economies, legalities, or societies. A user's distrust will not be lessened if we point out that the behaviour they fear can be controlled by a power they distrust!

5 Individual, Institutional, and Social Controls

We close this essay with a brief exploration of a taxonomic categorisation developed recently by Riegelsberger et al. [7]. After surveying a broad range of sociological enquiry, these authors propose a unifying framework. They consider one trusting relationship at a time. Accordingly, their framework has just two actors: a trustor and a trustee. The trustee has some power to harm or help the trustor, which we would represent in our framework by a directed arc from the trustee to the trustor.

The Riegelsberger framework defines trust as a three step process. In the first step, the trustor and the trustee exchange some information regarding the nature of the possible trusting relationship. Of particular interest in the framework are the "trustworthiness" signals from the trustee, which are dichotomised into symbols (pure information, uttered by the trustee) and symptoms (observations, either direct or indirect, by the trustor of the trustee's non-verbal behaviour).

In the second step, the trustor either withdraws or engages with the trustee.

In the third step, the trustee either fulfills the trustor's expectations, or it does not fulfill them. Riegelsberger's framework is focussed on the trustee's motivations for fulfilling or non-fulfilling. These motivations would be the result of security controls on the trustee in our framework, and fall into three categories.

The first category of security controls on the trustee's behaviour are *relational controls*. We do not adopt Riegelsberger's label of "temporal" for this category, because (as argued below) the temporal dimension of control in our framework dichotomises all three of Riegelsberger's categories. Relational controls arise because a trustor may maintain a record of their prior relations with the trustee. The retrospective form of this control is expressed if a trustor decides to withdraw because of their prior experience with the trustee. The prospective form of this control is expressed if the trustee's current decision is affected by the trustor's (presumed) record-keeping ability. The trustee may fear that the trustor will withdraw in the future, and the trustee may desire future engagements with the trustor. The commercial importance of relational controls is demonstrated by the prevalence of customer relation management (CRM) systems in modern enterprises. Gartner's estimated revenues for the global CRM market in 2008 was USD $9.15 billion.

Riegelsberger's second category are the *social controls* on a trustee. These controls arise because a trustor may be a member of a social group: a peerage in our framework. A peer may share their impressions of a trustee within the peerage. A trustee would thereby gain a reputation within the peerage for trustworthiness in certain respects, and they may also gain a reputation for untrustworthiness in some respects. The retrospective form of social control arises when the peerage defines and enforces (by solidary action) a normative control through their shared communications, mutual trust, and collective power to decide whether to engage with the trustee. The prospective form of this control arises when the trustor's behaviour is affected by their fear or desire for their relationship with the trustee's peers. The prospective form also arises when the peer group has formed a reputational estimate of the trustee, by a process analogous to the pricing of goods in an idealised free market, and when the prospective trustor uses this reputational information in their decision.

The third form of control on the trustworthiness of a trustee are the *institutional controls*. These controls arise when the trustee is subject to control by a hierarchy. The prospective form of these controls are architectural in nature: these arise when the trustee is effectively unable, due to a prior control exerted by the hierarchy, to refrain from fulfilling the trustor's expectations. In our framework, architectural controls may be the axioms of an intentional sys-

tem design, the unintentional consequences of an unaxiomatised design, or (in a model with theological axioms) laws of nature, random constraints arising in a randomly-defined system, or a god's decrees. The retrospective form of these controls arise when the trustee is subject to legal or regulatory sanctions. Any trustee who restricts their actions because of a behavioural trust in their hierarch's institutional control has transformed that hierarch's retrospective control into a prior restraint, i.e. an architectural control.

We conclude that the broad outlines, as described above, of Riegelsberger's framework are highly congruent with our framework. Before conducting this analysis, we were unaware of the importance of each actor's self-control arc in a security analysis, and our taxonomy of control did not clearly cover the individual controls. These can be considered a subcase of the architectural and legal controls, in that they are exerted by a hierarch who rules only himself. However we now believe individual controls are important enough, and easily enough to overlook, to deserve the following elaboration in our framework.

A trustor may have *individual trust* in their ability to engage with trustworthy trustees; this ability is dependent on their memory and judgement. A trustor may have *social trust* in the beneficial influence of their peerages' reputation system on trustees, and also in their ability to use the reputation system to select trustworthy trustees. Finally, a trustor may have *institutional trust* in the beneficial influence, on trustees, of a government, corporation, or any other system or organisation that can control (via rewards, punishments, enablements, and disablements) the trustee's actions.

Three types of distrust are also possible. An individual may mistrust their ability to avoid untrustworthy trustees, they may mistrust their peerages' reputation systems, and they may mistrust their legal systems and other hierarchical controls.

Our analysis of Riegelsberger's framework suggests the following hypothesis about social capital. Any sufficiently uniform population, i.e. a society such as a functional nation, will (we suspect) be biased in their decision-making toward considerations of individual, social, or institutional trust. Assessments of trust as a form of social capital may thus be misleading, unless these three types of trust are taken into consideration in the assessment. We are not conversant with the relevant literature, and would welcome pointers.

We briefly consider two limitations of Riegelsberger's framework. The three-step trust process does not include a dispute resolution mechanism, or any other explicit method for handling the cases in which a trustee's perception of a non-fulfillment disagrees with a trustor's perception of a fulfillment. Riegelsberger's framework is silent on this question, however dispute-avoidance and dispute-resolution issues are very important considerations in the practical design of trusted systems. In our framework, trustworthiness cannot be judged in any model that lacks axioms defining trustworthiness; and if the underlying ethic is deontological or consequentialist, then the axioms must establish the ground truth of a fulfillment. Anyone observing a real-world system may use a different set of axioms when constructing a model of that system; and this difference will

generally result in a different ground truth for each model. In a trustor-centric model, the ground truth of a trustor's trustworthiness might be established by an ethic, by polling the trustors, or by requiring every trusting relationship to have a trusted third party who, in case of dispute, establishes the ground truth regarding fulfullment. In a trustee-centric model, either an ethic or trusted third party is required.

Riegelsberger's framework is based on sociological research, so it is understandably sparse in its characterisation of technical requirements on designed systems. These have been studied extensively, see e.g. [8]. The "quality in use" concept of ISO 27000 is, essentially, a trustor-centric definition of trustworthiness ability that was developed by an intersubjective process involving hundreds (possibly thousands) of active contributors. A trustworthiness motivation may be demonstrated by an ISO 9000 accreditation.

6 Our Hopes

We hope that this essay will inspire technologists to consider a broader range of trust-enhancement strategies, when they attempt to develop trustworthy systems which actually inspire trust. We hope our essay will provoke sociologists, psychologists, and market researchers to consider how their theories and experimental results on trust and decision-making might be expressed in a sufficiently reduced fashion to be both understandable and useful to technologists. We also have a fond hope that our technologically-inspired, and highly reductionist, musings on the fundamental nature of trust will help to clarify future sociological or psychological studies.

References

1. Thomborson, C.: A framework for system security. In Stamp, M., Stavroulakis, P., eds.: Handbook of Information and Communication Security. Springer (2010) 3–20
2. Luhmann, N.: Trust and Power. Wiley (1979) English translation by Howard Davis et al.
3. Luhmann, N.: Familiarity, confidence, trust: Problems and alternatives. In Gambetta, D., ed.: Trust: Making and Breaking Cooperative Relations. Blackwell, New York (1988) 94–107
4. Parsons, T.: Suggestions for a sociological approach to the theory of organizations, part 1. Administrative Science Quarterly **1**(1) (1956) 63–85
5. Barber, B.: The Logic and Limits of Trust. Rutgers University Press (1983)
6. Lewis, J.D., Weigert, A.: Trust as a social reality. Social Forces **63**(4) (June 1985) 967–985
7. Riegelsberger, J., Sasse, M.A., McCarthy, J.D.: The mechanics of trust: A framework for research and design. International Journal of Human-Computer Studies **62**(3) (2005) 381–422
8. Côté, M.A., Suryn, W., Georgiadou, E.: In search for a widely applicable and accepted software quality model for software quality engineering. Software Quality Journal **15**(4) (2007) 401–416