A General Definition of Trust

Working paper, 5th August, 2012

Kieron O’Hara

Web and Internet Science
Electronics and Computer Science
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom
kmo@ecs.soton.ac.uk

Abstract: In this paper a definition and conceptual analysis of trust is given in terms of trustworthiness. Its focus will be as wide as possible, and will not be restricted to any particular type of trust. The aim is to show the key parameters that enable us to investigate and understand trust, thereby facilitating the development of systems, institutions and technologies to support, model or mimic trust. The paper will also show the strong connection between trust and trustworthiness, showing how the subjectivity of trust reveals itself in attitudes toward others’ trustworthiness; to trust someone/something is to believe that he/she/it is trustworthy. Both trust and trustworthiness are context-dependent, but the relevant contexts are different depending on whether one is trusting or trustworthy. Finally, the paper will discuss some of the complex issues connected with the alignment of trust with trustworthiness.

Introduction

In this paper I shall attempt to produce a general definition, or conceptual analysis, of trust. The focus will be as wide as possible; I do not want to restrict my analysis to any particular type of trust. I wish the analysis to apply to trust that is arguably hard-wired, and to trust that is entirely rationally-based, to any kind of appropriate agent (human, animal or artificial), and in the context not only of individuals but also institutions and organisations. The paradigm of a relationship involving trust is usually thought to be that between two individuals who are able to size each other up in imperfect but indicative ways. However, the evolution of the concept does not determine exactly how it can and should consistently and usefully be extended, and so although this paradigm is important, I will not take it as a typical case. Similarly, it is plausible that trust began as a moral concept, and it carries a great deal of ethical force; nevertheless, trust need not always be applied in morally-suffused situations, and I will not treat the moral dimension as definitive.

The aim of the analysis is to reveal the key parameters that enable us to investigate and understand trust, to work out how and why there can be too much trust in a society, or too little, and to develop institutions and technologies that facilitate well-placed trust, or mimic its effects in appropriate ways via design. Of course, such an analysis will not solve problems, but hopefully will at least enable problems to be stated accurately.

The analysis is of trust, and of the related idea of trustworthiness. I do not intend to delimit or define related ideas such as risk, complexity, uncertainty and confidence (although I shall make some brief remarks about reliance and reliability). It may be that the boundaries between these ideas are quite porous anyway, or that there is little consensus over exactly what the relationships are. In this paper, I also do not intend to critique other theories, or to contrast this theory with them; I will merely state a definition, and so this working paper contains no references.

Apart from this introduction and a conclusion, the paper has three substantive sections. The first discusses the important and prior concept of trustworthiness. The second then defines trust, and discusses the various components of the definition. The third discusses what is often formulated as the problem of trust, the difficulty of aligning trust and trustworthiness effectively and accurately. A short conclusion discusses future work.

Trustworthiness

The essential prior concept for understanding trust is *trustworthiness*. Trust is an attitude that one takes to the trustworthiness of another; in turn, the other’s trustworthiness is a property that they have. Broadly speaking, it is the property that they will do what they say they will do. If they fail, then it will typically be for some reason outside their control.

Trustworthiness is naturally not context-independent; one is not trustworthy in all respects in all contexts. One might be a trustworthy car mechanic without being a trustworthy brain surgeon. One may be morally impeccable but somewhat naïve, so that one is trustworthy in the custody of money, but not trustworthy in the custody of a cunning child.

Trustworthiness can be expressed as a quadruple, as in formula (1).

(1) Tw<Y,Z,R(A),C>
where Y and Z are agents, R is a representation of behaviour aimed at an
audience A, and C is a context.

This states that Y is trustworthy. Throughout this paper, I shall use Y and Z as variables for agents. In particular, for ease of exposition and to help defuse ambiguities, I shall usually refer to Y using feminine pronouns. By this, of course I do not intend to suggest that only women can be trustworthy.

(1) should therefore be read as: Y is willing, able and motivated to behave in such a way as to conform to R, to the benefit of members of A, in context C. The role of Z will be explained below. For ease of exposition, I shall write of Y being trustworthy; unless noted otherwise, I will be assuming R and C as given. In other words, I shall assume that there is some context for Y’s trustworthiness.

The importance of a claim being made about intentions, capacities and motivations

As noted, trustworthiness is not usually completely general; one is trustworthy in specific respects. It would be ridiculous to try to judge whether, say, President Obama was a trustworthy teacher of differential equations, as he has not represented himself as either able or willing to teach higher mathematics of any kind. A person is trustworthy when she does what she says she will do. Hence Y is trustworthy only in the context of a claim that she has the intention, capacity and motivation to constrain her behaviour in some way. R is the relevant representation of the behaviour that she has the intention, capacity and motivation to perform, the content of the relevant claim. Y might be judged untrustworthy if she is unwilling to conform to R, unable to conform to R, or lacking incentive to conform to R once it has been asserted to the relevant audience.

Z is the agent responsible for creating and disseminating the representation R of Y's intentions, capacities and motivations. It is essential, therefore, that Z has the authority to make such a claim about Y.

Context

The context C is some type of relevant restriction of the circumstances in which Y is claimed to be willing, able and motivated to conform to R. C might be a particular type of task (fixing cars, not brain surgery), or might be a set of circumstances delineated by a particular role. Y may be willing, able and motivated to conform to R during office hours, when she is employed in a particular role, but unwilling to work out of hours. If she is, say, a lawyer, then she may be a trustworthy provider of legal advice between the hours of 9.30 and 5.30, but makes no commitment to answer requests out of hours or at weekends.

C may be extremely precise, or alternatively very sketchily drawn. It may even be that Y is unable to describe the context exactly; she may rely on her (and others') abilities to recognise when the situation is out of context. We may have to track her behaviour, to see when she is willing, able and motivated to conform to R, in order to determine (some of) the limits to C. C may even be variable over time, although it cannot be totally arbitrary or subject to random change. If C was subject to arbitrary change, then Y could not be said to be trustworthy; formula (1) would trivially entail that Y is trustworthy when she is trustworthy and not otherwise.

Typically, the more commercial or contractual the relationship between Y and others is, the more precisely delineated C will be. As Y's relationships become less formal and based less on mutual gain, then C will generally become more open-ended.

Agency

Who might Y be? Y can be anyone who may perform a task, upon whom one might wish to rely. Y, if trustworthy, is willing, able and motivated to conform to R. Therefore a statement about trustworthiness implies further statements about Y's intentions, capacities and incentives. Y must therefore be the type of thing that could have intentions, capacities and incentives.

What about non-human agents? Is it possible for Y to be non-human? In this paper, I shall take a neutral stance about this issue. Much depends on whether we interpret terms like 'intention', 'capacity' and 'incentive' for non-human agents. Non-human agents have capacities; that seems straightforward. The issue of whether a non-human agent can properly be understood as trustworthy or not will then turn on whether one's philosophy of mind allows metaphorical talk about the intentions and incentives for them. Clearly, say, a piece of software has no intention to work in anyone's interest, although it may *be intended to* work in the interests of those who have paid for it. Once it has been paid for, then it is permanently 'on call' for its user. The question of the application of the term 'trustworthy' depends on how seriously one regards these inverted commas. It may be that we should speak not of trustworthy software, or trustworthy companies, but of 'trustworthy' software and 'trustworthy' companies. We might talk in terms of reliability rather than trustworthiness, and a suggestion is made as to how we could understand that distinction below. But we should not expect a hard and fast distinction between reliability and trustworthiness, as they share a number of conceptual features.

A non-human agent might be an animal (as in a trustworthy guard dog or racehorse, whose behaviour is reliable), a piece of technology (as in a trustworthy piece of software or a bridge over a river), or an organisation (as in a company that provides a particular type of service). As long as their behaviour can conform to a representation of a behavioural ideal, then that will provide a source of evidence that these non-human agents could in principle be counted as trustworthy. I will write as if it is not a category error for non-human agents as these to be called trustworthy; however, nothing in the argument will hang on this assumption, and if the reader believes that trustworthiness can only properly apply to humans, or only to adults (or only to some classes of non-human agent, such as animals), then that will be consistent with the definition. The definitions of trustworthiness and trust are independent of these questions in the philosophy of mind.

And who is the mysterious Z? This needs to be someone with the credentials to issue a representation of behaviour to which Y will conform. In the most common case, Y = Z – Y will represent *herself* as willing, able and motivated to behave in certain ways.

Yet this will not always be the case. Where Y occupies some role in a company, then Z is the company which, or the officer who, defines the role that Y occupies, and takes it upon itself to certify (via job interviews, performance monitoring and the potential for using sanctions) that Y is indeed willing, able and motivated to occupy that role in a trustworthy manner. If Y is a piece of software, then Z might be the designer or the company that sells the software, or whoever issues the specification for the software’s performance. If Y is a racehorse, then Z might be the trainer who has trained Y using his tried and trusted methods to run fast and not to throw its jockey; Z’s claim R will then take the form of an honest appraisal of the horse’s ability to the jockey, to the owner or to the racing press.

It is of course essential that Z has the authority to issue the representation. If a political commentator determines that President Obama is untrustworthy because he has failed to perform some action (invading Iran, say), then we have to ask whether that commentator has the credentials to claim that Obama was willing, able and motivated to perform that action.

When Y ≠ Z, there is scope for confusion about who has authority to make claims about whom. For example, in industry the practice of outsourcing services can lead to problems. A company might claim that an operative Y is trustworthy in some customer service role, but because it has outsourced the services to a service supplier or call centre, that company may not as a matter of fact be able to determine what Y is willing, able or motivated to do. The secondary supplier might be the only body able to do that, which could lead to problems of reputational damage if the primary and the secondary company (and Y) differ about her intentions, capacities and motivations.

Representation

The claim R is a representation of how Y should behave in the ideal. R needs to be disseminated by a Z with authority to do so, and Y’s behaviour (in context C) must conform to R. If it does not so conform, then Y or Z must be able to show that some unforeseen event prevented Y’s behaviour conforming with R. As with C, R might be open-ended or quite precise.

Its precision may be at any one of a number of levels. For instance, an MP or a congresswoman may represent herself very precisely with respect to her ends, as working for the benefit of her constituents in various respects (improving their economic position, reducing crime in their neighbourhood, or whatever), but may leave open the exact means she will pursue. A doctor represents herself very precisely as working to restore her patients to health, but she may give herself a lot of leeway in how she achieves that.

Alternatively, R may specify very precisely the exact steps that Y will take in response to various contexts. Y may have remarkably little discretion, and must perform her task in very specifically defined ways. This is less usual; in the most usual type of case, one expects a trustworthy person to behave in ways that are congruent to a particular set of interests, as with the MP or the doctor. Depending how much of R is left open to interpretation, there is more or less scope for dispute. Y is trustworthy when she is willing, able and motivated to behave in certain ways; that does not determine that she will succeed in all respects.

In many cases, R will represent Y's duties, or her responsibilities. In these cases, then moral judgments may be made about Y's trustworthiness.

In (1), R has a place for a variable A, which denotes an intended audience. Y's claim R will be aimed at a group of people or agents. A may contain everybody, when Y hopes to be trustworthy without restriction, but this is not generally the case. One might be trustworthy to one's family, to one's tribe or kin, to fellow nationals, to one's clients or paying customers. A racist may be willing only to be trustworthy to those with whom she shares skin colour.

This audience could be conceived as an aspect of the task context, but it is important enough to separate it out as an extra variable. Y simply does not want a relationship with, or to be beholden to, everyone without limit. Her promises will at least sometimes be precisely targeted to certain individuals or groups. A mother will behave as a mother only to her children. She may be *in loco parentis* to other people's children for brief and defined periods, but her trustworthiness as a mother extends unconditionally only to her children. Put another way, Y's trustworthiness is intended *only* to benefit members of A (Y will typically be neutral about whether non-members of A benefit from her trustworthiness, whether directly or indirectly). A soldier may fight to defend the homeland, although citizens of allied nations may also benefit from his endeavour; the soldier will not mind that, although he would not be properly motivated if those allies were *all* that he was fighting for.

The notion of audience gives us a way in to understanding the content of R. The members of A are those whom Y intended to benefit from her trustworthiness. R should therefore convey in what way Y is willing, able and motivated to behave (including subordinating her own interests narrowly conceived), in order to serve the interests of members of A.

R may be explicit or implicit; more usually it will have explicit and implicit aspects. It may be that the expectations of Y's behaviour are almost totally unwritten and implicit; being a 'good neighbour', for instance, implies a lot of things and raises a lot of expectations, but is extremely open-ended. On the other hand, when a professional represents herself as providing a particular service, she may be relatively explicit about ends and even means. As R becomes more explicit, it begins to resemble a contract; an R that is more implicit is correspondingly less contractual. Even a very explicit R may also contain implicit limits. If Y were a car mechanic, she might well represent herself explicitly as willing, able and motivated to fix cars of a certain make with a certain range of maladies within a certain period of time for a certain fee, but will not feel it necessary to add what is also true, that after the repair she will return the car to the owner.

R may also be inclusive or exclusive. We may understand R as determining a set of things, however precisely specified, that Y *must* do. An MP must act in the interests of her constituents as set out in her manifesto. Or alternatively, R may simply rule out things, however, precisely specified, that Y *must not* do. The MP must not enrich herself or her family, or her political party, via the public purse. In most cases, R will have inclusive and exclusive aspects, which in turn may be explicit or implicit. A car mechanic must (attempt to) fix the car; she must not keep it for herself.

R will generally degrade gracefully. As we approach, and cross, the borders of context C, it is unusual for someone to decide suddenly not to conform to R. Quite often, people will do their best in somewhat unfamiliar circumstances. A doctor who specialises in one kind of medicine will do her best when presented with a different kind of illness when the right kind of specialist in unavailable (e.g. on an aeroplane). She should not be held to the same kind of standards as the specialist, of course, but will generally try to conform to a less stringent version of R as best she can.

This is not always the case; in the UK, the pejorative term ‘jobsworth’ applies to a person who refuses to conform to R the moment that the border of C is crossed (someone who will not work even five minutes after the office closes, or who will never interpret a rule in a sympathetic way – a person who refuses to help because “it’s more than my job’s worth”). And a piece of trustworthy software should be expected to crash (even if it does not) if the conditions of its designed function do not obtain.

Z (or Y) need not be responsible for the *content* of R. Z simply needs to be able to assert with authority that Y’s behaviour will conform to R in C. R may be determined by social processes; for instance, a trustworthy neighbour, or trustworthy mother, does not define what constitutes ideal neighbourliness or motherhood herself. These are determined by norms in a community, and Y simply needs to signal that she is prepared to conform to them. In many cases, R is understood across a community, and Z is able to assert that Y will conform to such a pre-existing R. In still other cases, as we shall come to note below, R is negotiated between Z and a trustor to meet the trustor’s specific requirements.

On other occasions, Z (or Y) does generate the content of R. A company that employs Y in a particular role will often define that role. Perhaps the most usual case is that Z defines the detail of the content of R within a socially-generated template; we all know broadly speaking what a car mechanic is willing, able and motivated to do, but the garage that employs her will define the detail, such as what times of day this applies, or what types of car she is qualified to work on. In that case, most of R is a social construct, which Z (the garage) has adjusted slightly to conform to its terms of employment of Y.

Finally, of course one’s intentions, capacities and motivations can be represented in different ways, and so one can be trustworthy and untrustworthy at the same time, depending on which representation one has disseminated to which audience. If Y is a spy, she could represent her intentions, capacities and motivations to her home country as R, and represent her intentions, capacities and motivations to her foreign employers as R’. If her behaviour conformed to R’ instead of R – e.g., instead of guarding the secret plans with her life, she photographs them and gives them to a foreign agent – then she would be deemed untrustworthy by the authorities in her home country, but trustworthy by the authorities in her foreign employer. There is no contradiction in this.

Trustworthiness versus reliability

It may be that the root of the distinction between trustworthiness and reliability (and by extension the distinction between trust and reliance) can be traced to the role of R. Where R is very inflexible, specific, explicit and precise, it may be that reliability is the more appropriate quality of Y, rather than trustworthiness. Where R is flexible, unspecific, implicit and/or imprecise, this may be a situation where Y is a candidate for trustworthiness. Trustworthiness is likelier to degrade gracefully than reliability.

However that may be, elaborating the distinction is not the purpose of this paper, so I shall not explore this thought in much detail, except to point out that the distinction is not a sharp one, and so one should expect the two concepts to blur into each other. It may be that trustworthiness is a species of reliability, in which case everyone/thing that is trustworthy is *ipso facto* reliable, but even so there will be tricky borderline cases. We often talk perfectly intelligibly of trusting, as well as relying on, calculators, bridges or plastic corks in wine bottles; this is to be expected, as trust and reliance, and trustworthiness and reliability, share a good many logical properties.

Generalised trustworthiness

Trustworthiness, as in (1), is usually relative to a task or a context. However, we are prone to describe certain people as trustworthy as a general character trait. Given (1), the interpretation of general trustworthiness is fairly straightforward to specify.

(2) If Tw<Y,Z,R(A),C> whenever a Z with authority properly represents Y as being willing, able and motivated to conform to R to the benefit of members of audience A, then Y can be seen as generally trustworthy.

In other words, if Y is trustworthy in all (or most) specific contexts where she has a duty, or is claimed, to be trustworthy, then she is generally trustworthy.

So, for example, if we revisit the case where Y is a spy, we see that, though she is trustworthy from the point of view of her foreign employers, we could not seriously entertain the wider proposition that she was a trustworthy person generally.

Trust

In this section, I shall define trust on the basis of the prior concept of trustworthiness. Having given a definition, I shall discuss some of the issues raised by the definition in more detail.

Broadly speaking, trust is an attitude toward the trustworthiness of an individual. In short, if X trusts Y, then X has a positive view of Y’s trustworthiness. If we take an agent’s attitude toward another agent to be a belief about that agent, we get:

(3) X trusts Y $=_{df}$ X believes that Y is trustworthy
where X and Y are agents.

If 'belief' is not considered equivalent to 'attitude held by agents' then this pleasingly simple formulation of trust will have to be replaced by the more complex sentence of English 'X's attitude toward Y is that she is trustworthy'.

I will use X throughout this paper as a variable for a trustor, and Y throughout as a variable denoting a trustee. I shall disambiguate by using masculine pronouns to refer to X and feminine pronouns to refer to Y. Once more, the reader should avoid the elementary error of confusing grammar and sexuality, and should make no assumptions about real-world gender on the basis of these grammatical conventions.

As we have seen, 'Y is trustworthy' is a complex proposition anchored to a context, and so X's attitude toward it also has a complex representation anchored to a different context of relevance to X. Because of the difference between the contexts in which X makes his judgment, and in which Y envisages her claim for trustworthiness, it is not quite as straightforward as X holding that Tw<Y,Z,R(A),C>. Trust should be represented as a 6-place relation, in which a proposition about X's attitude is embedded, as follows.

(4) Tr<X,Y,Z,I(R[A],c),Deg,Warr>
where X, Y and Z are agents as before, and I(R[A],c), Deg and Warr are qualifiers of X's attitude to Y's trustworthiness to be described in more detail below.

If we are allowed to take 'belief' as a shorthand for X's attitude toward Y, this 6-place relation can be read as 'X believes that Y is trustworthy, on some account proposed by Z, which X takes as entailing I(R[A],c). He has a confidence Deg in his belief/attitude in Y's trustworthiness, and the belief/attitude is based on a warrant Warr.' In the remainder of this section, I will expand on the functions of these variables.

Trust as an attitude

Trust is an attitude, which brings it into the scope of the philosophy of mind. I do not have to solve, or even to have a view on, these issues. For instance, it makes no odds to the definition of trust whether one is a representationalist, a dispositionalist, a functionalist or an eliminativist about attitudes, as long as one believes that we have attitudes, and that embedded in formula (4) is the sort of thing we may have an attitude about. Although I am sceptical about the common philosophical view of belief as necessarily concerning propositional attitudes, (4) is surely consistent with that common view.

In the rest of this paper, I shall talk of people 'making trust judgments'. By this I mean they will be coming to trust or not trust someone or something else. I don't mean to imply they have a strong level of control over their cognitive processes. Hence those who are of the opinion that people have very little control over cognitive process should take this sort of phrase as loose, metaphorical talk only. It does not affect the definition or the argument.

Finally, we must not lose sight of one important effect of trust being an attitude of an agent, which is that X's perceptions are paramount. The facts of the matter (which are important for trustworthiness) count for less than X's perceptions of the situation.

Agency

X, Y and Z are as noted in (4): X is the trustor, Y the trustee, and Z is the person making the relevant authoritative claim about Y’s intentions, capacities and motivations. This is all seen through X’s eyes, as we are describing an attitude X has toward Y. In particular, it should be the case that Z is authorised to make claim R about Y; however, what is important for trust is that X believes that Z is so authorised. Conversely, even if Z is authorised, if X does not believe that he is, he is unlikely to trust Y on that basis. Again the reader should consider the usual caveats about the use of the term ‘belief’ as a shorthand for an attitude toward Z; it is simply easier to construct an English sentence around the verb ‘believe’.

Instantiations of the X, Y and Z variables

How do X, Y and Z relate to each other? X is to have an attitude toward Y, Y will be trustworthy if and only if she is willing, able and motivated to conform to R, and Z is the authorised person who makes the claim that Y is willing, able and motivated to conform to R. X, Y and Z could all be different people.

The most obvious potential equivalence, as noted above, is that Z could be identical to Y. Y can make her own claims about her intentions, abilities and motivations, and more often than not this is the case. But other equivalences are possible too.

Can X be identical to Y? Yes, one can trust oneself – or, perhaps more strikingly, one can fail to trust oneself. A recovering alcoholic, for instance, might take his dog for a walk on a route that does not involve his passing a pub, because he does not trust himself to resist the temptation. Normally, one trusts oneself routinely, implicitly and ‘invisibly’, but in the case of the alcoholic, the day when he trusts himself to go past the pub without entering may be a big day in his life.

If X is identical to Y, then in most of those cases, they will also be identical to Z – in other words, the trustor, trustee and the guarantor of the trustee’s behaviour are all the same person. But it is also possible that trustor X and trustee Y be identical, while the guarantor Z is someone else. In the alcoholic example, the representation of the alcoholic’s good behaviour could be provided by a therapist (“you are cured, you will be able to pass the pub without entering”). Hence X trusts himself, possibly reluctantly, to behave as Z, the therapist, has assured him he will. To be sure, X, the alcoholic, also trusts the therapist Z – a recursive aspect of trust that I shall discuss in more detail below. But the main thing from this point of view is that X trusts X to conform to a representation of his intentions, abilities and motivations provided by a *different* person Z.

Could X be identical to Z without being identical to Y? In such situation, the trustor would be the person making the claim about the trustee’s behaviour. One apparent possibility would be that X is a manager who employs Y and dictates the representation of trustworthy behaviour that Y is expected to conform to in her job. X trusts Y to do the job, which X has also defined. However, this is not a case where the trustor is identical to the trustee’s guarantor; this would confuse two separate instances of trust. X trusts Y’s representation of herself as someone willing and able to do the job; here Y is being her own Z, i.e. representing her own intentions, capacities and motivations to X. However, a customer of X’s company, call him X_1, approaches Y as a member of the company, whose prospectus advertises that all its employees are willing, able and motivated to conform to R (defined by X). Hence in this second case, X is the person making claims about the trustee Y, but is not the trustor. At no stage does X believe *his own* claims about Y’s intentions, capacities and motivations. In the first case, X believes Y’s claims about her own intentions, and in the second he himself makes a claim about Y’s intentions that is believed by the customer X_1.

A better type of example is where X trusts Y without encouragement from Y, and supplies the representation R of Y’s intentions, abilities and motivations himself. For example, X may love Y, and persuade himself, without any help from Y, that she is willing, able and motivated to stay faithful and true to him. It is, perhaps, unlike that such trust will be well-placed, as X is not authorised to speak for Y. However, he may also persuade himself that he knows Y better than she knows herself, and that therefore he can represent her intentions, capacities and motivations accurately. Inappropriate trust of a person who is both unfaithful and not hypocritical about it is a common staple of fiction, as for example with Amelia’s love of George Osborne in *Vanity Fair*, or Rose’s love of Pinkie in *Brighton Rock*. At worst, such misplaced trust leads to the extreme and unmotivated feelings of betrayal that often lie behind the crime of ‘stalking’.

Another example of this type of trust relationship is where trust is ‘forced’ upon an untrustworthy person in order to socialise them. For example, it is a common strategy to give young people responsibilities in order to build their character, as with the Outward Bound educational organisation, which puts its charges into situations which creates in the individual a state of dissonance requiring adaptive coping, which leads to a sense of mastery when equilibrium is managed (as they note in their literature). The point here is that the claim (R) about the abilities of the learner is made by the educator (X), not the learner (Y) herself. Hence the educator not only takes the X role in the trust relation, but also the Z role.

In these cases where X = Z, X believes that he knows Y’s intentions, capacities and motivations better than Y does herself. The truth of X’s belief is not relevant for establishing the fact of X’s trust.

The role of roles

Trust and trustworthiness can be focused on individuals or on roles; I have already written briefly about the roles that can be played by a trustworthy person. When X trusts Y, for example to be his bank manager, then it may be that he trusts her because of her personal qualities, with which he is acquainted, or it may be that he trusts her because she occupies a role whose occupant he trusts. If Y moves on, to be replaced by Y_1, in the latter case, X will now trust Y_1, the new occupant of the role. In the former case, we can deduce nothing about X’s attitudes in the new circumstances. We can say nothing about his attitudes towards Y_1 as we have no evidence, and we can say nothing about his attitudes towards Y, because we do not know what new claims about Y’s intentions, capacities and motivations are being made since she moved on from her job (X may now believe that Y has been promoted above her capacities, for example).

A trustor can also find that he has to tailor his judgments to a role he is playing. A judge or a jury member should, if possible, suppress many of the trust judgments they make as an individual when they are in the courtroom. It may be that Y’s eyes being too close together make her appear untrustworthy, but that is not relevant evidence in assessing whether a crime has been committed by her. In any role with public responsibility, someone who was disposed not to trust people of a certain ethnic group would be very wrong to allow that disposition to interfere with his professional judgments. A football referee is well aware *qua* human being that a trivial contact between two players, trained athletes both, is extremely unlike to send one of them crashing to the ground in extreme agony, but *qua* football referee he is bound to treat the incident as one where physical contact has constituted foul play and where a stretcher is required to ferry the wounded hero from the ground in case there has been serious injury.

In each case, as we unpack the complex proposition that Tr<X,Y,Z,I(R[A],c),Deg,Warr>, the way that X, Y and Z are identified may well change the meaning of the expression. 'X trusts Y' (in some context) may have a different truth value depending on how the variables are filled in: 'John Brown trusts Mary Green' is different from 'John Brown trusts the Governor of Clink Prison', which is different again from 'Her Majesty's Inspector of Prisons trusts the Governor of Clink Prison'. The last, unlike the first, conveys nothing whatsoever about the personal relationship between John Brown and Mary Green.

Types of agent

X must be an agent capable of holding a complex attitude toward Y's intentions, capacities and motivations. Once more, questions of what type of agent that might be devolves to the philosophy of mind. Once more, I have no great problem with babies, non-human animals, artificial agents or organisations holding such attitudes, but the line might be drawn more tightly on philosophical grounds without doing harm to the definition.

I shall on occasion use terms such as 'belief' – this I hope will not entail that trust can only be held by humans. It may *be* the case that trust can only be held by humans, but that should be a separate argument from the logical structure of trust which I set out here. Hence my definition should be, and is, consistent with the view that non-humans and undeveloped humans can trust, and also be consistent with the view that trust is restricted to the human realm. If my occasional use of the term 'belief' is regarded as prejudicing this question, I hope the reader will give me a certain expositional latitude, and make the requisite substitution in his or her head.

It might also be the case that certain agents, even if capable of trust, are capable only of a circumscribed type of trust judgment. A baby, for example, might be deemed capable of trusting its mother in several open-ended and implicit respects, but quite out of its depth when it came to assessing which of several candidate solicitors would be most trustworthy for the purchase of a new house. A piece of software designed by a bank to assess whether someone applying for a loan is trustworthy in respect of paying back the loan on time would be hopeless for determining whether someone is a trustworthy car mechanic. Indeed, the software would be hopeless at determining the trustworthiness of the would-be borrower if the evidence it received diverged from the expected input for its program (e.g. it would be unable to make even a guess at the candidate borrower's trustworthiness based on signs such as the firmness of her handshake or her ability to look it openly in the eye, or evidence of her wealth such as her clothes, car and house). So a capacity to trust in some contexts would not imply a capacity to trust elsewhere.

The subjective elements

Trust, like trustworthiness, is context-dependent. As an attitude toward the trustworthiness of another it inherits the context-dependence of trustworthiness itself, but it brings in further subjective elements of its own. These are the final three elements of formula (4), which will be described in detail in this subsection.

The interpretation of commitments

If Y is trustworthy, then a claim R(A) must have been made about her intentions, capacities and motivations with respect to an audience A in some context C. When X trusts Y, then he must interpret that claim. R, to recall, is often, perhaps usually, implicit; even if explicit, there will be complex or borderline cases, and it may involve statements about Y's interests or intentions, rather than exact specifications of behaviour. So it is not a trivial matter for a potential trustor to work out how Y's intention to conform to R in C will result in particular behaviour that is desirable for X in the contexts that X is interested in.

Hence X's trust in Y will include only an oblique reference to R in its description. The fourth term in (4) is written I(R[A],c), in which c is the particular context or set of contexts in which X has his own specific interests. To be properly applicable, it should be that $c \subseteq C$, otherwise of course the claim about Y's intentions, capacities and motivations does not apply. Furthermore, it must also be the case that $X \in A$, otherwise it will not be Y's intention that X gain from her trustworthiness; if $X \notin A$, Y promises X nothing.

Assuming $c \subseteq C$, then X will interpret R in c. In other words, he will form expectations of how R will circumscribe or dictate Y's behaviour to his own benefit (assuming $X \in A$) in c, which I have written as I(R[A],c), taking I as a function on $R(A) \times C$.

X must believe that $X \in A$ and $c \subseteq C$, and that R(A) entails I(R[A],c) in the restricted class of contexts c. That does not mean that these propositions are actually true, or that Y or Z believes they are.

It follows from that that it is not necessarily the case that I(R[A],c) correctly describes what Y will do, but that does not necessarily mean that Y has let X down. For instance, a theme in Sherlock Holmes stories is that Holmes behaves in ways which surprise and dismay his clients. They are cross that he is thinking about 'trivial' matters, and are ignoring what seem to be the important factors. Of course, Holmes is proved right in the end; he was trustworthy all along (he was working in his clients' interests, as he claimed he was), but the clients' interpretation of that claim was inaccurate. In general, the more explicit that R is, presumably the more likely it is that I(R[A],c) is accurate, although there will always be room for a mismatch. On the other hand, if R is implicit but relies on commonly shared social norms (such as well-understood notions of financial probity), then I(R[A],c) may be spot on.

Hence I(R[A],c) introduces three important subjective elements of trust – first of all, the *interpretation* of the claim about Y's intentions, capacities and motivations; secondly, the restriction of the application of Y's trustworthiness to *the range of contexts which interest* X; and thirdly, the belief that X *is part of Y's intended audience*.

Note finally that the situation need not be such that X is presented with a *fait accompli* by Z. The content of R, the membership of A, or the meaning of I(R[A],c), can be (and often is) negotiated between X and Z – X sets out his requirements for interaction with Y, and Z crafts a set of incentives or directions for Y that meet X’s requirements. That may reduce the possibility of mismatch between trustor and trustee, but of course there is always scope for misinterpretation and failure to agree on the basic semantics (as well as deliberate subversion by Y).

To summarise the import of the first four variables of (4), we should read it as expressing that X believes that Y’s intentions, capacities and motivations conform to I(R[A],c), which X also believes is entailed by R(A), a claim about how Y will pursue the interests of members of A, made about Y by a suitably authorised Z.

Degree of confidence

The fifth variable in (4) is Deg, which is the degree to which X believes that Y is trustworthy, or the confidence that X has in his attitude. The intuition here is that one has stronger attitudes about some things than others, and that some attitudes are more firmly or confidently held than others. This is an important parameter in the analysis of trust for two reasons.

First of all, we are able to make comparative judgments about trust. It is important in many respects to know that X trusts Y_1 *more than* Y_2. For example, it may explain why he bought an item from Y_1 rather than an identical but slightly cheaper item from Y_2. In that particular case, it would allow a rational explanation of X’s purchasing decision. All things being equal if X’s degree of confidence in Y_1’s trustworthiness is higher than his degree of confidence in Y_2’s trustworthiness, X is more likely to rely on Y_1 than Y_2.

Secondly, trust is an important risk management tool, and Deg helps explain risk management strategies. Broadly speaking, if X trusts Y, he will be willing to risk some of his assets in transactions which could be affected by Y. The greater his confidence in Y’s trustworthiness, all things being equal, the greater the value of the assets he will be prepared to risk (the greater the risks he will be prepared to take).

Degree of confidence is therefore an important parameter. How do we measure it? The definition I am proposing is neutral about this, so once again one could apply to a number of disciplines depending on one’s purposes. One could analyse real-world data (for example, from an online auction site) about genuine decisions to find a metric. One could perform experiments in the lab, perhaps within a game-theoretic framework. One could simply sketch a model, maybe modelling degree of confidence in a fine-grained way as a real number between 0 and 1, or on a qualitative scale that could be quite coarse. One could aim for psychological realism in the trust judgments that resulted, or for justifiable rationality (for example, in some kind of automated trust advisory system). My own view is that nothing very complex is required as a description of judgments in daily life (maybe a 5-point qualitative scale enabling broad comparisons and coarse-grained risk judgments), but nothing in the theory outlined in this paper hangs on that intuition.

We can now summarise the import of the first five variables of (4). We should read it as expressing that X believes, with confidence Deg, that Y’s intentions, capacities and motivations conform to I(R[A],c), which X also believes is entailed by R(A), an authorised claim about how Y will pursue the interests of members of A, made about Y by a suitably authorised Z.

Warrant

The final variable in (4) is Warr, which stands for the warrant for X’s belief, attitude or judgment about Y’s trustworthiness, and also the warrant for the particular value of Deg that is associated with it. In other words, it describes the positive and negative input to X’s judgment.

Once more, (4) is neutral over what type of thing Warr might be. As long as Warr explains the judgment, then it is adequate. It is worth emphasising that Warr need not be a set of propositions, and the trust judgment should not assumed to be rational (although of course sometimes it will be). It should also be insisted that the warrant is intended to explain the trust judgment, not to persuade others of its validity.

This is an area for research in logic, sociology, psychology and neuroscience, and no doubt many more disciplines. Examples for factors that Warr might cover include:

- Propositions from which X has derived a trust judgment, whether rationally or irrationally (for instance, after research into various service providers).
- A statement or model of Y’s reputation.
- X’s memory about past dealings with Y.
- X’s views of the role Y plays.
- Further trust judgments (perhaps X trusts Y’s employer).
- Recommendations of Y by others.
- Y’s qualifications or credentials.
- A credible and costly commitment made by Y (for instance, if X is not satisfied, Y will give him his money back).
- Sanctions X can apply if Y defaults.
- A response to subconscious signalling (for instance, people with features such as a symmetrical face tend to be trusted more, while people with other features such as facial hair tend to be trusted less.
- A response to conscious signalling (for instance, Y is wearing a suit, or smiles a lot).
- A hard-wired neuropsychological process (for instance, a baby trusting its mother).
- Responses to various chemicals (for instance, doses of oxytocin can increase trust).
- Peer pressure.
- Sexual attraction, or general ‘liking the look of’.
- Feelings of racial or gender solidarity.

Trust can be manipulated, and can be deeply irrational. It can also be highly rationally placed, and human society has developed a number of institutions, both formal and informal, which facilitate accurate trust judgments. Hence any general definition of trust such as (4) should be neutral between the sensible and the idiotic judgments. Nevertheless, it seems plausible to suggest that there should be *some* warrant, however misguided, for X’s trust in Y. Note also that even some of the more rational judgments X makes can be subverted if Y is devious enough; for example, qualifications can be faked, recommender systems spammed, and reputations manipulated.

The warrant for a judgment is an important part of predicting what the repercussions of the judgment will be. For instance, how X will act if Y turns out to be untrustworthy depends at least partly on the ground for his original judgment.

We can now summarise the import of (4). We should read it as:

> (5) X believes, with confidence Deg on the basis of warrant Warr, that Y's intentions, capacities and motivations conform to I(R[A],c), which X also believes is entailed by R(A), a claim about how Y will pursue the interests of members of A, made about Y by a suitably authorised Z.

Trust as an action: placing trust

Trust is an attitude or belief, and so – unless one is a behaviourist of quite a reductionist degree – need not be manifested constantly in behaviour. In particular, the oft-made connection between trust and risk, while real, is not a necessary or internal connection.

We should distinguish between X *trusting* Y, and X *placing trust in* Y. In the former case, (4) is true for X and Y. Given the truth of (4), X can start to place trust in Y by acting in various ways that would appear irrational otherwise.

> (6) X places trust in Y $=_{df}$ X performs some action which introduces a vulnerability for X, and which is inexplicable without the truth of (4)

If (4) is true, then X trusts Y and he does not therefore believe that his action in placing trust in Y *has* introduced much of a vulnerability, because he believes that Y's behaviour will conform to I(R[A],c) as explained above. Note that trusting someone is conceptually prior to placing trust in them, but that one can trust without placing trust.

There are other ways in which trust can be evidenced without X placing any trust or exposing himself to any vulnerabilities. Most obviously, one can perform a simple speech act: X can announce "I trust Y". This puts him at no risk and introduces no vulnerability. A similar type of behaviour is to make a recommendation, for example on eBay or by pressing a 'like' button about an Amazon review.

Recursive trust; grounding trust

As hinted earlier, trust is not grounded; it may be based on further trust. So, for example, if X trusts Y to conform to some claim about her behaviour made by Z, part of the warrant for that might be that X trusts Z when the latter makes a supposedly authoritative claim about Y's behaviour. As an example, Y may be a bank manager whom X trusts because he trusts the bank. His trust in Y does not ground out in other attitudes or beliefs – in this case it depends on further trust in Y's institution.

Put in the terms of proposition (4) above, the Warr variable may contain further instances of trust to support the top level claim.

To take an example, the warrant for X's trust in Y may be based on Y's reputation for plain and fair dealing. If X has not dealt with Y before, then his knowledge of her reputation will rest on some external source such as the confidence rating on eBay, for instance. In that case, the strength of the warrant for X’s trust in Y will be (partly) dependent on the strength of his trust in the eBay rating system.

Whether trust can be grounded totally – i.e. whether a trust claim could be made which does not feature another trust claim as part of its warrant – is a moot point. However, we should be aware that trust in something will often be based on trust in something else. Trust in an online bank may be based (in part) on trust in the https protocol, which in turn may be based on trust of the browser and operating system of one’s computer, which may in turn be based on trust of one’s brother-in-law who has an IT qualification and who installed the system, which in turn may be based on trust of one’s sister (which may or may not be hard-wired). However that may be, we cannot expect such trails of dependent trust judgments to always bottom out, even if our patience in tracing them might.

The problem of trust

Trust has many benefits for society – it makes interaction easier – and so it is often argued that levels of trust should be raised. This is not the case – if we trust people who are not trustworthy, then interactions will become more risky and costlier. We cannot think about trust without considering trustworthiness at the same time. The purpose of this section is to sketch the fundamental political and social requirements that trust and trustworthiness pose.

Costs and benefits: a prisoner-style dilemma

There are many tragedies in the human condition, but one of them is the following. The trustor X benefits from the trustworthiness of the trustee Y; unfortunately he controls only his own trust. The trustee Y benefits from being trusted by X; unfortunately she controls only her own trustworthiness. The result is a prisoner-style dilemma when two agents come to cooperate in such a way as to require trust. The situation looks like this:

X\Y	**Trustworthy**	**Not trustworthy**
Trusts	Maximum benefits of cooperation	X loses, and Y gains, whatever he risks
Does not trust	Y loses whatever she invests in her trustworthiness	No interaction, so neither benefit nor harm. Opportunity costs for both agents.

Let us plug in some numbers. Let us assume that X and Y may collaborate on a task such that, if it succeeds, they each gain $10. To succeed, X must risk some assets, let us say $3, and must trust Y to deal with them fairly. In turn, to be trustworthy is not cost-free for Y; she must make certain security arrangements which will cost her $1 as a sunk cost. An example would be a bet on a horse. Y can get information about a sure-fire cert which will cost her in hospitality for her informant, but has no extra money for a bet. X has money for a bet but no inside information. The obvious arrangement is that Y obtains the information, at some cost to her, from her informant, and X gives her the money for the bet. They agree to split the winnings equally. X, if he does not trust Y, can simply not fund the bet. Y, if she is not trustworthy, will not bother obtaining the information from her informant, but instead will pocket X’s stake money if she can. The matrix now looks like this:

X\Y	**Trustworthy**	**Not trustworthy**
Trusts	10\9	-3\3
Does not trust	0\-1	0\0

Looking at the matrix, X might reason that if he trusts, he may gain $10, but equally could end up $3 down. If he does not trust, he will at least be square whatever happens. Not trusting may well look like the safe option. Similarly, Y might reason that if she is trustworthy, she may be $9 in front, but equally she could have spent a dollar for no return. On the other hand, if she is not trustworthy, she can’t end up worse off than she is, and may pocket $3.

Hence trust and trustworthiness pose a coordination problem that will be hard to solve using only rational self-interest as a means.

Stating the problem

The problem is simple to state, and of course it is not that we merely should wish to raise levels of trust, or to create a high-trust society. The main issue is to try to ensure that *we trust all and only trustworthy agents*, or at least to approach that state as an asymptote. There should be a causal connection between the trustworthiness of one agent and the formation of an attitude of trust towards her by another.

Similarly, rather than trying to spread trust, we should aim for *a high-trustworthiness society, in which trust is placed effectively*, thereby reducing both the possibility of fraud and the opportunity cost of failing to trust appropriately, while still maximising the opportunities for cooperation and interaction.

The twofold nature of the problem, the need to connect an attitude with a property, creates the well-known moral hazard of trust. A would-be trustee should be trustworthy, and take steps to ensure that her actions always fall within the authoritative claims being made about her. This can be quite an investment in resources, and *even that* will not necessarily lead to her being trusted. To get trust, she will also need to market herself in appealing ways, to get the message through to potential trustors. The danger is always that she may reason that if trustworthiness is not enough to get trust, and if marketing will make the difference, why should she bother with trustworthiness at all? Maybe marketing will be enough.

Tactics

Dealing with the problem of trust therefore means establishing the causal connections that increase the probability that a trust judgment one way or another is correctly placed. There are a number of ways of doing this, of which the main four are:

- Signalling. Would-be trustees send signals of their trustworthiness to potential trustors (the intended audience A). These signals may be consciously-crafted, such as wearing a suit or smiling agreeable, or unconscious, such as maintaining eye contact or not sweating. They may come in more socialised or explicit forms, such as via reputation, endorsements or credentials. One important class of signal is something that is a by-product of a successful operation; for instance, a trustworthy professional will often earn a lot of money, and therefore her fashionable clothes and expensive car are telling signals.

- Reciprocity. Very often trust situations involve reciprocal agreements, perhaps implicit, between trustor and trustee. People offer services and expect other services in return, whether or not these are laid down in a formal or semi-formal contract. Reciprocity need not take the form of a pair of coordinated actions; it may simply be the expectation that X can rely on Y in some circumstances, and that Y will return the favour in others (for example, each agrees to babysit for the other's children in principle, without setting up a series of dates when this will happen). In such an arrangement, it may be that only one part of the bargain is ever fulfilled (for example, because as a matter of fact Y's oldest daughter has reached an age where she is able to babysit for her siblings, and so Y never actually needs a babysitter), but this does not count as a defection or reneging by X and Y will not hold it against him.
- Institutions. Society sets up institutions that enable people to trust each other. Indeed, trustor and trustee may never even have any idea of their mutual involvement, as with a bank when savers lend to borrowers, but mediated through the bank itself. The bank takes upon itself the effort of investigation of borrowers, thereby lowering the cost and increasing the effectiveness of such investigation, through economies of scale. Other institutions, such as contract (in those societies where contracts are routinely respected), enable trustor and trustee to deal directly with far more confidence than two strangers would otherwise.
- Sanctions. If a potential trustor has available to him a suite of effective sanctions that he can apply to a would-be trustee, then his confidence that the trustee will behave in a trustworthy manner will, all things being equal, increase. He will be able to convince himself that the trustee is more likely to judge that trustworthiness is in her own interests, as it would be also in her own interests to avoid sanctions. One advantage of most contract systems is that in most countries it also puts a set of (legal) sanctions in the hands of the trustor (however, sanctions are not necessary to enable contracts to function, if in a culture people generally respected them).

This is certainly not an exhaustive list, but signals, reciprocity, institutions and sanctions are important mechanisms. None of them is foolproof. To take the most obvious example, if a signal can easily be counterfeited, then a would-be trustee can send the signal without actually being trustworthy. Similarly, institutions do not 'solve' the issue of trust, they merely shift its focus. The trustor's trust in the trustee becomes dependent upon his trust in the institution, with the potential for systemic collapse. If a trustee proves untrustworthy, then the trustor may cease to trust the institution that brokered the deal. But his trust in the institution may be the only, or the chief, warrant for his trust in many other trustees – which he may in consequence withdraw *en masse*. The effectiveness of reciprocity can depend on whether the reciprocal acts are performed simultaneously; if there is a lag in time then it is always possible for the second trustee to renege on her obligation. Finally, the usefulness of sanctions depends entirely on how painful they are for the trustee, how easy they are for the trustor to apply, and how easily they may be avoided. An untrustworthy trustee will no doubt spend effort and time working out ways of working with the system for her own ends – for example, sticking to the letter but not the spirit of the law.

The direction of causality

As to which mechanisms to establish causal links between trust and trustworthiness will work, much depends on the direction of causality between the two. Perhaps the most common view is that trustworthiness causes trust. Y goes around being trustworthy for a period, and gradually earns X’s trust. If she defects (or more to the point is perceived by X to have defected), then she will lose X’s trust. The onus in that sort of transaction is on Y to ensure that she send the right signals, ideally backed up with trustworthy behaviour. X is seen as making a judgment on prior behaviour, including Y’s reputation.

If this is taken as the canonical direction of causality, then there are two obvious problems. The first is how to bootstrap trust. If X has little or no information about Y, he cannot place his trust in her. In a wider sense, if a new community develops – perhaps online, or perhaps around a new market for a new type of good or service – then X has no information about *any* of the participants, and so it is hard to see how trust will develop between any pair of actors in such circumstances.

The second problem is how someone who has (or is perceived to have) transgressed can regain trust. Does she start from zero, or less than zero? Reputational damage is extremely corrosive on this model of causality.

However, there is another direction of causality, which is from trust to trustworthiness (it is an indication of how far game theory has influenced our psychological models that many would claim this is counterintuitive). X trusts Y, and from that experience Y becomes trustworthy. She learns, via X’s input, what constitutes trustworthy behaviour in the relevant contexts, and becomes socialised. This model renders X far less of a passive judge and more of an active participant in shaping the understanding with Y. It also makes Y less of a supplicant and more of a negotiator, working with X to determine the standards of her future behaviour.

Of course, in reality the causal relations between trust and trustworthiness go in both directions, depending on the situation. Trustworthiness-to-trust is perhaps more characteristic of business/contractual/Gesellschaft relationships, while trust-to-trustworthiness is more likely to be found in more social/Gemeinschaft relationships. But there is no clear dividing line between the two, and there are no hard and fact ‘rules’ about which direction is better.

Furthermore, the waters are muddied in the general case by the existence of other types of relation. Most obviously, we often trust, or otherwise, on the basis of qualities that are obviously unconnected with trustworthiness. We tend to trust people with symmetrical faces, and are less likely to trust people with facial hair or whose eyes are too close together. These extraneous factors complicate the story still further.

Conclusion

This working paper has defined trust as a belief in someone’s trustworthiness, and delved in some depth into the conceptual analysis of these two terms. This is only a beginning of the correct description of trust, and further issues need to be examined. Firstly, we need to examine the different kinds of failure of trust, and how we can respond to these. Second, we need to highlight and consider different types of trust – exactly how and on what grounds trust is placed in concrete social situations. A final question is the social role of trust in managing complex interactions with others. These questions will be addressed in future versions of this paper.