The Alan Turing Institute

Technical Briefing

Risk Models of National Identity Systems: A Conceptual Model of Trust and Trustworthiness

Authors

Dr Paul Smart, Dr Brian Pickering, Professor Michael Boniface, and Professor Dame Wendy Hall

UNIVERSITY OF Southampton

The Institute is named in honour of Alan Turing, whose pioneering work in theoretical and applied mathematics, engineering and computing is considered to have laid the foundations for modern-day data science and artificial intelligence. It was established in 2015 by five founding universities and became the United Kingdom's (UK) National Institute for Data Science and Artificial Intelligence. Today, the Turing brings together academics from 13 of the UK's leading universities and hosts visiting fellows and researchers from many international centres of academic excellence. The Turing also liaises with public bodies and is supported by collaborations with major organisations.

The Alan Turing Institute

British Library

96 Euston Road

London

NW1 2DB

Table of Contents

Introduction

The present report summarizes the work undertaken in respect of the effort to develop a conceptual model that supports the trust-related evaluation of National Identity Systems (NISs). This report was written as part of the Risk Models of National Identity Systems (RM-NIS) project, which forms part of the Trustworthy Digital Infrastructure for Identity Systems initiative coordinated by the Alan Turing Institute.[1] The general aim of the RM-NIS project is to provide a framework that informs the development and evaluation of NISs from a trust perspective. In particular, the RM-NIS project seeks to support the trust-related evaluation of NISs from a multi-stakeholder perspective.

In addition to describing some of the more general issues associated with the attempt to model trust-related concepts, the present report outlines a parametric approach to modelling trust and trustworthiness. This is intended to circumvent some of the definitional problems that have confronted previous attempts to subject trust-related concepts to analytic scrutiny. The present report also seeks to advance our understanding of the relationship between trust, uncertainty, and risk. It does this by directing attention to some of the cognitive processes that are relevant to the formation of trust-related cognitions and the implementation of trust-related actions.

In addition to discussing the results of recent analytic and modelling efforts, the present report highlights a number of issues and concerns that will be the focus of future work. These include the modelling of system-specific features that are relevant to the trust-related evaluation of NISs. We also discuss some of the issues raised by a multi-stakeholder approach to the trust-related evaluation of NISs.

1 Understanding Trust and Trustworthiness

In order to make progress in modelling trust and trustworthiness, it helps to have a basic understanding of what the terms “trust” and “trustworthiness” actually mean. This is one of the major stumbling blocks in trust-related research, for there is, at the present time, no consensus on the precise meaning of these terms. At a general level, theoretical approaches to trust can be divided into so-called doxastic (or cognitive) accounts of trust (e.g., Hardin, 2002) and affective (or emotional) (Jones, 1996; Lahno, 2020) accounts of trust. Of these two basic types of account, doxastic accounts tend to be more popular within the philosophical, scientific, and engineering literatures. It is, indeed, a doxastic account that best describes our own approach to the modelling of trust-related concepts.

According to doxastic accounts of trust, trust entails a belief about the object of trust, where

[1] https://www.turing.ac.uk/research/research-projects/trustworthy-digital-infrastructure-identity-systems.

the object of trust is typically referred to as the trustee. Let us denote the trustee using the symbol Y and the object or entity that trusts the trustee (also known as the trustor) as X. According to doxastic accounts of trust, the statement "X trusts Y" entails that the trustor (X) possesses a belief about some property of the trustee (Y), specifically, the belief that Y is trustworthy. These trust-related beliefs are what we will call *trust beliefs*. Relative to the present account, the concept of trust can be understood solely in terms of trust beliefs. That is to say, the term "trust" is applied to situations in which we ascribe a trust belief to a trustor as a means of explaining (or predicting) that trustor's attitude towards some trustee. The thing that distinguishes trust beliefs from other kinds of belief is the content of the belief, or the thing the belief is about. We suggest that trust beliefs are about the trustworthiness of a trustee. Trust thus refers to a state-of-affairs in which some X (trustor) believes that some Y (trustee) is trustworthy. To say that "X trusts Y" is equivalent to saying that "X believes that Y is trustworthy." This echoes the views of prominent trust theorists, such as Hardin (2002) who also endorses a doxastic approach to trust.[2]

It should be clear from this characterization that trust and trustworthiness are not the same. In line with a number of other theoretical accounts of trust, we suggest that trust is an attitude that we take towards objects that we deem to be trustworthy, and trustworthiness is a property of the thing that is trusted (McLeod, 2020; O'Hara, 2012). It should also be clear that by casting trust as a particular kind of belief (i.e., a belief about the trustworthiness of a trustee), much of the definitional burden associated with a theoretical account of trust is shifted from the notion of trust to the notion of trustworthiness. If trust is a belief, then we already know what trust is: trust is a particular kind of belief. What makes trust beliefs special is the content of these beliefs, i.e., the thing that trust beliefs are about. Given that trust beliefs are about a particular kind of thing— namely trustworthiness—then we can understand something about the distinctive nature of trust beliefs by understanding something about the peculiar nature of trustworthiness. This sort of idea is consistent with prior attempts to define the trust concept. As noted by O'Hara (2012):

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 thatthey will do what they say they will do. If they fail, then it will typically be for some reason outside their control. (O'Hara, 2012, p. 2)

[2] To be a little more specific, we are embracing what is called a pure doxastic account of trust. According to Keren (2014; 2020), pure doxastic accounts are to be contrasted with impure doxastic accounts. While pure doxastic accounts view trust as nothing more than the possession of trust beliefs, impure doxastic accounts emphasize that trust beliefs are not sufficient for trust.

To understand, trust, then, we first need to understand the notion of trustworthiness. In particular, we need to understand what it means for a trustee to be trustworthy. Unfortunately, this is where we run into a further definitional quandary, for different theorists have characterized the notion of trustworthiness in different ways. Many trust theorists agree that the notion of trustworthiness is tied up with the idea that a given object will behave in a particular way. The point of disagreement relates to what it means for an object to behave in a particular way that is relevant to the fulfilment (or betrayal) of trust (Goldberg, 2020). Consider, for example, that we might expect an inanimate object, such as an alarm clock, to behave in a particular way. When we set the alarm, we expect the alarm to sound at the appropriate time and thus wake us up. In this sense, we are relying on the alarm clock to do something, and the reason why we rely on the alarm clock is due to our expectations about how the alarm clock will behave. Despite this appeal to reliance and expectation, it isn't clear that the notion of trust is the best way of explaining why we rely on the alarm clock. If the alarm fails to sound, then we would surely be wrong to suggest that the alarm clock had betrayed us. We would, no doubt, be disappointed by the alarm clock's sonic shortcomings, but it isn't clear that we ought to regard this as anything to do with the inaccuracy of our beliefs about the clock's trustworthiness. We might have believed that the clock was reliable when, in fact, it was not; but it is commonly assumed that reliability-related beliefs are not the same as trust beliefs. As noted by Hardin (2002), we can rely on a great many things to operate in certain ways, but mere reliability does not seem sufficient for trust. Indeed, Hardin (2002) goes as far as to suggest that the notion of trust makes no sense in fully deterministic settings; i.e., in settings where the behaviour of the trustee is determined by physical laws and/or overly rigid incentives:

> I do not trust the sun to rise each day, at least not in any meaningful sense beyond merely having great confidence that it will do so. Similarly, I would not, in our usual sense, trust a fully programmed automaton, even if it were programmed to discover andattempt to serve my interests—although I might come to rely heavily on it. (Hardin, 2002, p. 12)

As we will see, a number of efforts to understand the notion of trustworthiness have appealed to the idea that trustees must possess certain properties (see Section 8). At a minimum, trustees must possess the ability to do what they are trusted to do, although this is typically deemed to be insufficient for trust, at least in situations where the trustee is a human individual. In addition to ability, it is commonly assumed that the trustee must be suitably motivated to do what they are trusted to do, such that they will *choose* to pursue a particular course of action in preference to other action alternatives.

A natural question to ask at this point is whether trustworthiness is a fixed property that can be quantified independently of any given trustor. One of the immediate problems confronting

this idea is that the same trustee may be judged to be more or less trustworthy by different trustors. Perhaps, however, this is merely a consequence of the fact that different trustors perceive the same trustee in slightly different ways, thereby leading to divergent views as to the trustee's trustworthiness.

Unfortunately, there is a deeper problem with the idea that trustworthiness can be understood in some trustor-independent fashion. The nature of this problem emerges as soon as one considers the way in which the interests of one group of trustors are sometimes diametrically opposed to another group of trustors. In situations where the interests of two parties are in conflict, Y cannot be equally trustworthy to both parties. In general, Y's trustworthiness is specific to certain individuals or collections thereof. Y may be trustworthy to some individuals or groups, but not others. In general, a trustee cannot be the trusty servant of two masters whose interests are in conflict with each other. The members of a military platoon may trust one another with their lives, but it would be foolish for an enemy combatant to regard a member of the platoon as trustworthy simply because he is deemed to be trustworthy by his comrades. The same goes for any future arsenal of 'trustworthy' military robots, drones, and other autonomous combat systems. These systems may be trusted by those who control them, and they may indeed be trustworthy relative to those who control them. They are not, however, uniformly trustworthy: They serve the interests of one particular group of combatants, and they do so at the expense of another group of combatants. There is, as such, no such thing as a uniformly trustworthy drone, any more than there is a uniformly trustworthy soldier. The trustworthiness of these things is relativized to particular trustors (and/or groups of such trustors). It is, indeed, the untrustworthiness of a trustee to one group of trustors that underwrites its trustworthiness to another group of trustors.

The importance of this particular point should be relatively clear when we direct our attention to technological systems and devices. We often talk of trustworthy systems as if there was something about such systems that makes them uniformly trustworthy. But issues of trustworthiness are tied up with the things that a trustee might be expected to do and the sorts of trustors that might come to rely on them. Just because a technological system is trustworthy to one group of trustors, this does not mean that it is trustworthy to *all* potential trustors. In the effort to design trustworthy systems, it is easy to overlook this relational aspect to trust. It is, of course, important to understand what it is that makes a system trustworthy, and how we might go about building trustworthy systems. But before we embark on this engineering effort, we ought to ask ourselves whose interests such systems intended to serve? Who, in particular, are the trustors that stand to benefit from the trustee's trustworthiness? We might be inclined to think that technological systems ought to be trustworthy to as wide a group of people as possible, perhaps the class of human individuals (the members of the global human

population). This, however, is unlikely to be realistic, especially given the fractious nature of human relationships at a variety of social scales. If a trustee cannot be equally trustworthy to all people all of the time, then the interests of some people will need to be prioritized at the expense of others. But who are these people, exactly? And on what basis do we distinguish between those who will benefit from the introduction of technological and social interventions and those who will not?

A doxastic account of trust highlights the importance of trustees to our attempts to make sense of trust-related behaviour. In particular, a doxastic account of trust suggests that in our efforts to understand and model trust, we ought to consider the way in which trust beliefs are directed to particular trustees. While there have been some attempts to study trust as a property of individual trustors, the general consensus is that trust can only be understood once we have determined who or what is being trusted. This looks to be largely uncontroversial, for a question such as “Do you trust?” really makes no sense unless we specify some target object whose trustworthiness is to be evaluated. When it comes to NISs, for example, we assume that trustors will form beliefs about the trustworthiness of NISs, and these beliefs will be relevant to the differential behaviour that is exhibited towards these systems. In choosing one system over another, for example, a stakeholder may regard their decision as being motivated by issues of trustworthiness. There may, of course, be multiple reasons why a given stakeholder opts to choose one system (A) over another system (B), but trustworthiness is, at least one of the factors driving this choice. If we ask the stakeholder why they chose A rather than B, they may respond by saying “I trust A more than B.” From the standpoint of a doxastic account of trust, this response is equivalent to saying “I chose A over B because I believe A is more trustworthy than B.”

At a minimum, then, it seems to appropriate to regard trust as being relativized to particular trustees. This is typically what trust theorists mean then they suggest that trust ought to be considered as a “relational phenomenon” (see Cook & Gerbasi, 2009). In addition to trustees, however, there is good reason to think that trust ought to be conceptualized as a three-place relation of the form X trusts Y to φ, where φ refers to some sphere or domain of activity (A. Baier, 1986; Hardin, 2001, 2002). This claim makes a great deal of sense when it comes to trustees with whom we are unfamiliar. We might, for example, trust a neighbour to look after our pet hamster while we are on holiday, but we wouldn’t necessarily trust them to look after our children. The same is true in even the most intimate of social relationships. Robbins (2016), for example, notes that:

I may…trust my wife, but not surely for anything and everything. I might trust her to edit a paper or maintain fidelity but not with medical advice or to fly a plane. Under these conditions, I might assume that her motivations toward my interests are the same,regardless of the matter

at hand, but her ability to actualize each of these matters will vary. As a result, my beliefs about her trustworthiness will vary from matter to matter. (Robbins, 2016, p. 978)

In view of this, it appears likely that trust is always relativized to particular areas, spheres or domains of activity. This is not to say that this sort of φ-related scoping of trust relationships is always made explicit in our everyday discourses about trust. In common parlance, we often refer to individuals as being trustworthy, or we say that we trust them without qualifying this by saying "with respect to φ." Nevertheless, it should by now be clear that trust beliefs are always relativized to spheres of activity, even if this detail is hidden by the elliptical nature of vernacular use of trust.

The relativistic nature of trust entails a number of methodological implications. In particular, it raises doubts about the methodological adequacy of certain techniques that have been used to explore trust. Historically, much of the work on trust has relied on the use of survey techniques. These surveys typically ask the respondent to report on the extent to which they trust particular entities or types of entities. Consider, for example, the following question, which is taken from the General Social Survey (GSS)—a survey administered by the National Opinion Research Center (see Hardin, 2002, p. 201):

Generally speaking, would you say that most people can be trusted, or that you can't betoo careful dealing with people?

Aside from the fact that the respondent is being asked to make a judgement about a rather nebulous social group (namely, "people"), this question lacks any sort of reference to a particular sphere of activity. What, exactly, are people being trusted to do? We might trust the majority of people not to knife us to death in the street, but would we be content to entrust the care of our children to the majority of the national population, most of whom, of course, we know nothing about. The GSS question is too vague to yield anything in the way of a useful response. If we interpret the question to mean "random strangers that I might trust with the care of my children," then we are likely to give a very different response to someone who interprets the question as meaning "random strangers that might mug me at knife point."

Also note that the way in which respondents interpret the meaning of the term "most people" is apt to be a further source of confusion. What does "most people" really mean in this situation? Does it mean people who I might meet, but currently do not know, or does it mean people that form part of my inner social circle? If it is the latter, then responses are likely to be heavily biased in favour of a positive trust response. If you ask me whether I can trust my friends, then I am likely to say "yes," for what would be the point in preserving friendships with people who I do not trust? My friends are trustworthy precisely because they are my friends. If they were not trustworthy (with respect to at least some matters), then I would not count

them as my friends.

A consideration of φ is of particular importance when it comes to a consideration of technological systems. For the most part, technological systems are designed to do a particular thing, and their abilities are accordingly scoped to particular areas of activities or to particular kinds of functionality. Inasmuch as a technological system is reliable, we may trust it to do the thing that it is designed to do, but beyond this activity-related context, the behaviour of the system is unpredictable, and trust diminishes. In general, we can only gauge the trustworthiness of technological systems relative to the kinds of activities they were designed to perform. This is what is sometimes referred to as the zone of trust, especially when it comes to research into the trustworthiness of Artificial Intelligence (AI) systems (see Grodzinsky, Miller, & Wolf, 2020).

At first sight, this discussion about the relativistic nature of trust might seem unimportant when it comes to a consideration of NISs. One reason to think this is that we are concerned with a particular kind of trustee whose trustworthiness is being evaluated relative to a given domain of activity. The values of the X, Y, and φ variables are thus relatively easy to specify once our attention is directed to a particular kind of evaluative context. Suppose, for example, that we are interested in trust evaluations pertaining to one or more NISs. The trustor (X) is, let us suppose, a human individual who needs to make a decision about whether or not to adopt a particular NIS. The NIS, in this case, is the object whose trustworthiness is being evaluated, and it therefore qualifies as the trustee (Y). The value of φ will then be the activities that we expect the NIS to perform.

Perhaps, not surprisingly, this somewhat simplistic formulation hides a degree of complexity. Leaving aside the fact that X and Y may take on a number of different values, the appeal to activities (note the plural form) suggests that we might evaluate the same system in different ways depending on the specific kinds of activity or functionality that are important to us. A given NIS may thus be deemed to be trustworthy relative to its ability to support the reliable identification of individual citizens, but it may be deemed much less trustworthy relative to its ability to protect the privacy of those citizens. Inasmuch as different stakeholders direct their attention to disparate features of the same system, then they may arrive at very different views as to the more general trustworthiness of a given NIS. This sort of function- or feature-specific evaluation of technological systems is pretty commonplace. Consider, for example, the way in which we might judiciously moderate our trust in online systems based on the particular kinds of functionality they provide. Smart and Clowes (2021) suggest that Google search can be regarded as both a conventional search engine—providing pointers to online resources in response to specific queries—and as a question/answering system that delivers factual responses to specific questions. From an epistemic standpoint, they suggest that we can trust

the informational deliverances of Google Search in its capacity as a question/answering system, but we ought to be a little more circumspect when it comes to the content returned by the search engine. From a trust perspective, then, it becomes a little hard to know what we ought to say about the trustworthiness of Google Search, for this sort of claim can only be evaluated once we consider the way in which the same basic system is used for particular epistemic purposes.

In summarizing our work to advance our understanding of trust and trustworthiness, we suggest that trust is best understood from a doxastic, and more specifically, from a folk psychological perspective. In particular, we suggest that the "term" trust refers to a doxastic state (i.e., a state of belief) whose content is the trustworthiness of a given trustee (a trustee being the object or a trust relation and the target of trust evaluation). One of the virtues of this way of thinking about trust is that it ties our conceptual understanding to the folk psychological apparatus of thought ascription, and, perhaps more importantly, the role that folk psychological constructs play in the explanation of behaviour. This highlights the importance of explanatory concerns in our attempt to understand the notion of trust. In particular, we suggest that trust (qua belief) is a folk psychological explanatory construct that is invoked in particular situations as a means of making our behaviours (or behavioural propensities) intelligible to both ourselves and others. What is important here is emphasis on "particular situations." This is perhaps the best way of understanding the trust concept. We understand trust, not by directing attention to the properties of the trust qua concept, but to the properties of the situations in which the explanatory appeal to trust is warranted. Those situations are ones in which we voluntarily choose to make ourselves vulnerable to the actions of another object in the absence of any sort of guarantee about whether that object will 'choose' to act in the way we want them to (and, to a lesser extent, whether the future situation will allow that object to act in the way we want them to act). These situations are a common feature of our interpersonal interactions and exchanges due to the vagaries of human behaviour and the various 'hidden' causal forces and factors that drive such behaviour. Beyond these contexts, the term "trust (qua explanatory construct) is invoked in situations where we have some sort of analogical resonance to these paradigmatic trust situations, i.e., situations in which the behaviour of a trustee is governed by an inter-animated nexus of causal forces and factors whose machinations are seldom amenable to direct observation. (In this sense, the causal forces and factors amount to hidden or latent variables.) All of this fits extremely well with a so-called free energy approach to cognition, which see the computational imperative of the biological brain as one of reducing an information-theoretic isomorph of statistical free energy (Clark, 2013, 2016; Friston, 2009, 2010; Hohwy, 2013). It also accords well with recent work in deep machine learning where the aim is to identify the deeply-nested hidden (or latent) causal forces

and factors that shape the statistical structure of training corpora (Hinton, 2010; Smart, in press).

2 Trust Types

The trust literature is dominated by a discussion of different kinds of trust. Some examples include the likes of interpersonal trust, institutional trust, online trust, social trust, generalized trust, and so on. One of the advantages of conceptualizing trust as a three-place relation, of the form X trusts Y to φ, is that it provides us with a set of parameters that can be used to make sense of these types, kinds, or forms of trust.

Consider, first, the different forms of trust that emerge once we direct our attention to the nature of Y, i.e., the type of entity whose trustworthiness is the target of X's trust beliefs. A consideration of the type of Y, yields the following forms of trust:

Interpersonal Trust: the sort of trust we encounter in human interpersonal contexts. (Y = another human individual).

Institutional Trust: Y = institution.

Organizational Trust: Y = organization. As in, X trusts Microsoft.

Categorical Trust: Y = a type rather than a token/instance. As in, X trusts commercial organizations.

System Trust: Y = a technological or socio-technical system. [This is one of the forms of trust that is likely to be relevant to NISs.]

Self Trust: X = Y. Trust in oneself.

If we restrict Y to the set of human individuals, and then generalize over Y, we arrive at what is called generalized trust (sometimes called social trust). Generalized trust is thus the sort of trust we have over a collection of entities of a given type (in this case, human individuals).

Other types of trust emerge once we direct our attention to the nature of X:

Collective Trust: X = a collective entity, such as an organization or a nation state. One might try to explain peaceful Anglo-American relations by saying that the UK and the US trust one another. By contrast, one might explain the rather fraught status relationship of Anglo-Iranian relations by saying that they do not trust each other. In both these cases, X and Y are collective entities.

Public Trust: This can be interpreted as the average trust expressed by members of some

group towards some trustee. For example, public trust in democratic institutions was undermined by the actions of the President (given the status of Y as a political institution, this also qualifies as a form of institutional trust).

Finally, trust may be considered from the standpoint of φ. Robbins (2016), for example, distinguishes between so-called simplex trust and multiplex trust. Simplex trust is the form of trust that applies to a specific φ or a limited set of φs. Multiplex trust, by contrast, is the form of trust that is encountered in situations where φ concerns a large array of activities; for example, all those activities that we could reasonably expect a typical human being to perform in ecologically-normal circumstances.

The aim of the present report is not to survey the various kinds of trust that have been discussed in the trust literature, and we have thus refrained from an exhaustive survey of actual (and/or possible) trust types. Despite this, it should be clear that a consideration of trust parameters does provide the basis for an improved understanding of the trust-related intellectual terrain. In particular, it provides us with a relatively straightforward means of taxonomizing trust.

3 A Parametric Approach to Trustworthiness

In Section 2, we encountered some of the problems associated with the effort to define the notion of trustworthiness. While the definitional effort remains a prominent focus of research attention, it is unclear whether this sort of effort will terminate in a successful resolution of the definitional problem. One possibility is that trustworthiness is something that can only be understood once we limit our attention to specific trust-related contexts. Inasmuch as this is the case, then our understanding of trustworthiness would vary on a case-by-case basis. What it means for a human individual to be trustworthy, for example, might be very different from what it means for a NIS to be trustworthy. If this were to be the case, then the prospects for an all-encompassing definition of trustworthiness would start to look a little dim.

As an alternative to the definitional effort, we suggest that scientific efforts to study trustworthiness may benefit from a consideration of the general features of trust relationships. This is what we will call a parametric approach to trust. We have already encountered an example of this sort of approach, for the attempt to understand trust as three-place relation of the form X trusts X to φ is one that features an appeal to parameters (i.e., X, Y , and φ) that are deemed to be applicable to all trust relationships. This approach, it should be clear, is not committed to a particular theoretical account of trust or trustworthiness, in the sense that it

does not tell us what trust is or what it means for something to count as trustworthy. Nevertheless, the provision of these three parameters does yield an intellectual payoff. In Section 2, for example, we saw how a consideration of the relational and relativistic nature of trust imposed constraints on the practical effort to study trust via the use of survey techniques. Similarly, in Section 3, we saw how a consideration of trust parameters could support the theoretical effort to taxonomize the different forms of trust that have been discussed in the trust literature.

In previous work, we have sought to apply a parametric approach to the modelling of both trust and trustworthiness (O'Hara, 2012). In the present section, we provide a parametric characterization of trustworthiness and discuss how each of these parameters might be applied to NISs. In the subsequent section, we expand this analysis to include the notion of trust.

From a parametric perspective, we suggest that trustworthiness (Tw) can be modelled as a four- place relation of the form:

Tw<Y,Z,R(A),C> (1)

Each of these parameters will be discussed in greater detail below. For present purposes, however, it should be noted that Y and Z are agents, R is a representation of behaviour aimed at an audience A, and C is a context.

The parameter Y was introduced in the previous section. It refers to the trustee whose trustworthiness is to be evaluated. (1) is a claim about the trustworthiness of Y. It should be interpreted as the claim that 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" (O'Hara, 2012, p. 2). This way of characterizing trustworthiness is neutral as regards the nature of Y. Y is, in short, any object to which the notion of trustworthiness might be applied. Some theorists have sought to impose constraints on the sort of objects that can qualify as trustee, to the effect that something can only qualify as a trustee if it possesses goals, motivations, intentions, agency, volition, and so on.[3] For present purposes, however, we do not impose any constraints on the sort of objects that might be regarded as *bona fide* members of the class of trustees; a trustee is simply an object that is perceived to be trustworthy to a greater or lesser extent by a trustor.

The parameter R corresponds to a representation of how Y will behave in some situation. More specifically, R is a representation of how the members of A can expect Y to behave. In short, R is a vehicle for communicating information about what we earlier referred to as φ. No claim is made about the physical properties of R. R could, for example, consist of a written

[3] With regard to the possibility of institutional trust, for example, Hardin (2013) argues that one cannot really trust an institution because institutions cannot truly care or intend—only the persons within institutions can do that.

statement, or it could be some form of verbal agreement. R can also vary according to its scope. With regard to NISs, for example, R could represent one aspect of the larger functionality of a NIS or it could refer to the general properties of the NIS. Finally, R can vary according to its precision. The behaviour of a NIS may thus be specified at a very general level, or, alternatively, its behaviour may be represented in exquisite detail.

The goal of R is to specify what the members of A can expect from Y in situations where trust is to be fulfilled. A, recall, represents an audience, and R is intended to be communicated to the members of this audience. In general, the constituents of A can be regarded as the agents who are the actual or potential trustors of Y. They are the agents that can legitimately expect their trust in Y to be fulfilled should they opt to place their trust in Y. They are, in short, the individuals who are intended to benefit from Y 's trustworthiness.

The notion of an audience is intended to capture the idea that trustworthy trustees are selectively trustworthy to particular trustors. Trustworthiness, recall, is relative to particular trustors, and not every individual who might be in a position to form beliefs about Y 's trustworthiness can necessarily assume that Y will be trustworthy (to them). Having said this, A may be specified quite broadly. In the extreme case, a technological system may be designed so as to be trustworthy to all the human inhabitants of Planet Earth, in which case all human individuals (at least those resident on Planet Earth) would be members of A. At the other extreme, A may be limited to a particular group of individuals, or even a specific human individual. (Imagine, for the sake of illustration, a state-of-affairs in which I am trustworthy to my wife but no one else. In this case, A would be constituted by my wife, and my wife would be the only trustor that could legitimately trust me.) At this point, it should be clear that part of the challenge, from the trustor's perspective is to determine whether or not they are a member of A. If they believe that they are a member of A, then they are at least the potential beneficiary of Y 's trustworthiness. If the trustor should fall outside this group, however, then trust judgements and decisions are much more fraught. Inasmuch as R makes it clear who the intended audience is, then those who are not included in A can avoid disappointment by refraining from failing to place their trust in Y.

One way of understanding the appeal to R and A in (1) is via the notion of commitments. Commitments are important, in the sense that they make it clear what the trustor can expect of the trustee, and they also help to single out those occasions in which the trustee's behaviour ought to be seen as relevant to issues of trustworthiness. Suppose, for the sake of argument, that X trusts Y to φ, but Y is unaware that X is relying on them to φ. If Y is unaware that X is relying on them in some way, then it is hard to see how Y could be deemed to have betrayed the trust that was placed in them. Indeed, in the absence of any sort of understanding about how X might be affected by Y 's actions, then it is hard to see how Y could be deemed to be

an appropriate target of a trust relationship. Y 's trustworthiness is not impugned by their failure to do what X expects them to do if Y has no knowledge of how their actions might affect X.

Commitments help to resolve this ambiguity. If X is a member of A and Y makes it clear that the members of A can rely on Y to φ should the members of A place their trust in Y, then X has legitimate grounds for disappointment should Y fail to deliver on their commitments. If Y is genuinely trustworthy, then they should only commit to doing those things that they can deliver on. If a system fails to do this, then it risks being perceived as dishonest or inept at assessing its own abilities, neither of which are conducive to perceptions of trustworthiness. Conversely, if a system only commits to doing the things that it can do, and it always lives up to its commitments, then, the system's willingness to commit to φ is a reliable signal that it can be trusted to φ.

It might be thought that the primary beneficiaries of commitments are trustors. Providing a trustee is honest, sincere, and suitably selective about its commitments, then the trustor can use commitments as a guide for trust-related decisions. If, for example, a trustee commits to φ, then the trustor knows that the trustee will attempt to fulfil the trust that is placed in them. This is something that clearly benefits the trustor; but there are reasons to think that the trustee also derives some benefit from commitments. If X trusts Y to φ, but Y should be unable to fulfil the trust that is placed in them, then Y 's failure to φ might be interpreted as a failure of trustworthiness, with all the social, financial, and reputational damage that attends such failures. Given that the costs of failing to fulfil trust may be worse than the disappointment that comes with refusing a commitment, it may be of considerable benefit to the trustee to avoid a state-of-affairs in which their actions are misinterpreted as violations of trust.

As noted by a number of theorists, there is often a burden that comes with trust relationships, and commitments serve as a way of making it clear what can and cannot be expected of the trustee (Hawley, 2014). For the most part, trust is deemed to be something akin to a valued commodity in the trust literature, but that does not mean that trust is always welcome or desirable. At the very least, the trustee has an interest in limiting the extent to which their behaviour is subject to normative scrutiny. Sometimes we are content for our behaviour to be transformed into what amounts as a test of trustworthiness, but this is not always so.

In (1), Z is an agent with the authority to issue a representation of behaviour to which Y will conform. Z need not be separate to Y. Indeed, in a great many cases, Y = Z, and Y will thus represent themselves as willing, able, and motivated to behave in a manner that conforms to R. In other cases, Z will be an agent distinct from Y. This will typically be the case, when Y is a non-human individual, such as a NIS. In this case, Z may be the designer or company responsible for the production of Y. Z may also be a certification authority or social institution

that provides some assurance about Y 's abilities. It is even possible to imagine situations in which Z might be a computational system that reliably communicates information about Y. Suppose, for example, that Y = a NIS and a trust evaluation system is deployed to report on the system's ability, functionality, reliability, security, and so on. In this case, the trust evaluation system is communicating information about Y 's propensity to behave in a way that is consistent with the trustor's expectations, and it is therefore performing a role similar to that of a certification authority.

The final parameter to be discussed is C, which represents the circumstances in which Y is claimed to be willing, able, and motivated to perform the φ specified by R. The inclusion of C draws attention to the way in which trustworthiness is apt to change as one moves from one context to another. In general, the trustworthiness of a trustee is limited to a particular region of space and time. In respect of time, for example, it is important to note that trustworthiness is seldom unaffected by the passage of time. A technological system or platform that performs perfectly well today, may not perform so well in the future. Technologies may become obsolete as the result of new innovations or other shifts in the technological landscape. In addition, new cyber-security threats may emerge that threaten to transform a previously secure system into one that is burdened with vulnerabilities. In highly dynamic environments, trustworthiness may be confined to narrow temporal windows, and trustees may need to be subject to more or less constant evaluation. At the very least, those engaged in the evaluation of a system's trustworthiness are required to monitor the evolving situation and assess when their beliefs about a trustee's trustworthiness may be called into question.

The notion of context also includes assumptions about the situation in which a system is expected to operate. Systems may perform perfectly well in some environments, but not in others. In addition, there may be background conditions that are required to ensure the successful performance of a system. Hardly any trustee, human or otherwise, is unaffected by a shift in the situation in which they are embedded. A human individual may be extremely trustworthy when it comes to the completion of an important programming task, but if a power outage should occur, and the individual's access to computational resources should be curtailed, then the programming task will not be completed on schedule. Whether this particular failing ought to be chalked up as a failure of trustworthiness is, of course, unclear, since the individual did not necessarily choose to cease their programming activity; nevertheless, a consideration of the context in which trust fulfilment actions are to be performed is of vital importance when it comes to decisions about whether or not one should place their trust in a trustee. Situational concerns are also important when it comes to NISs. NISs, as with all national-level systems, are implemented within a given socio-political context, but this context is not fixed in perpetuity. In countries with democratically elected governments,

the values and priorities of national governments can sometimes shift in the wake of elections, and this can influence the extent to which a previously trusted system continues to be regarded as trustworthy.

4 A Parametric Approach to Trust

In the previous section, we saw how a parametric approach might be applied to trustworthiness. In this section, we attempt to apply the same approach to trust. Whereas trustworthiness was characterized as a four-place relation, trust (Tr) is characterized by a six-place relation:

Tr<X,Y,Z,I(R[A],c),Deg,Warr> (2)

Here, X, Y, and Z correspond to agents, and I(R[A],c), Deg, and Warr are qualifiers that are applied to the belief about Y 's trustworthiness. X, recall, refers to the trustor—the agent who evaluates the trustworthiness of Y and subsequently forms a belief about Y's trustworthiness. Given our commitment to a doxastic account of trust, this six-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). X has a confidence Deg in their belief about Y's trustworthiness, and the belief is based on a warrant Warr."

In view of the parametric approach to trustworthiness presented in Section 4, if Y is trustworthy, then a claim R(A) must have been made about Y's motivations, abilities and intentions to φ with respect to a particular audience A in some context C. When X is called upon to evaluate Y's trustworthiness, X must interpret this claim. If it should be the case that $X \notin A$, then Y makes no commitment to X, and thus there is no guarantee about Y's trustworthiness relative to X. In this sense, it should be clear that one of the things that needs to be determined is whether X is a member of A. In general, it must be the case that $X \in A$, otherwise it will not be Y's intention that X benefit from Y's trustworthiness.

X must also determine that the context (c) in which their own interests are likely to be satisfied is subsumed within the context in which Y's trustworthiness is assured (i.e., C). Accordingly, to be properly applicable, it should be the case that $c \subseteq C$, otherwise the claim about Y's intentions, capacities and motivations does not apply.

Assuming $c \subseteq C$, then X will interpret R in c. In other words, X will form expectations about Y's behaviour relative to X's interests. This subjective assessment of the extent to which X believes that $c \subseteq C$ and $X \in A$ is denoted as I(R[A],c) in (2). In fact, I(R[A],c) introduces three important subjective elements of trust. Firstly, there is the interpretation of the claim about Y's

motivations, abilities, and intentions; secondly, there is the restriction of the application of Y's trustworthiness to the range of contexts that are relevant to X; and thirdly, there is the belief that X is part of Y's intended audience.

The Deg parameter in (2) refers to the degree to which X believes that Y is trustworthy, or, alternatively, the confidence that Y has in their belief about Y's trustworthiness. Deg is informed by a number of factors, many of which relate to the reliability or completeness of the information that is available to support assessments of Y's trustworthiness. We seldom have perfect information against which to gauge the trustworthiness of a trustee. Indeed, if such information were to be available, then it is unclear whether we would need to invoke the notion of trust in explaining/justifying our actions towards a trustee. Giddens (1990, p. 33), for example, suggests that:

There would be no need to trust anyone whose activities were continually visible and whose thought processes were transparent, or to trust any system whose workings were wholly known and understood. It has been said that trust is 'a device for coping with the freedom of others,' but the prime condition of requirements for trust is not lack of power but lack of full information. (Giddens, 1990, p. 33)

One reason to think that trust beliefs ought to be associated with some measure of confidence or certainty is because trust-related decisions are often motivated by comparative judgements of trustworthiness. Suppose that X is required to choose between two trustees (Y_1 and Y_2), both of whom are identical and thus of equal trustworthiness. If X has access to more reliable information about Y_1 as compared to Y_2 then (all things being equal), X is likely to have greater confidence in their beliefs about Y_1's trustworthiness than they are Y_2's trustworthiness. Accordingly, X may decide to place their trust in Y_1 in preference to Y_2 based on the greater confidence they have in Y_1's trustworthiness.

The final parameter to be discussed is Warr. Warr refers to the warrant for X's belief about Y's trustworthiness, as well as the Deg that is associated with this belief. In short, Warr refers to the reasons (both positive and negative) for X's beliefs about Y's trustworthiness. It subsumes all the information that influences the process by which X's trust beliefs are formed. This could include, for example, X's prior experience of interacting with Y, or trustees similar to Y. (This is what is sometimes referred to as experience-based trust). It could also refer to reputational information about Y. If Y has a good reputation, and X has reason to believe this information is credible, then X may believe that Y is trustworthy in the absence of any direct information about Y. (This is what is sometimes referred to as reputational trust.)

5 Trust, Risk, and Uncertainty

When X trusts Y to φ they typically expect Y to φ. This does not mean, however, that they know for certain that Y will φ. Trust situations invariably involve some degree of uncertainty. We can never know for certain that Y will φ, even though we might strongly expect them to φ. Such uncertainty leads to risk. In general, when we trust someone, we want them to act in a particular way, and the reason we want them to act in a particular way is that we derive some benefit from them acting in that way. Conversely, if the trustee should fail to act in the way we expect them to act, then we incur a cost. This is what makes trust decisions risky: We want to be sure that our exposure to risk is minimized, and we do so by seeking to understand the extent to which a trustee will do what we expect them to do. But given that trust situations typically involve some sort of temporal delay (see Coleman, 1990), and we are thus (for the most part) attempting to predict what might happen in some future state-of-affairs, then there is always an element of risk associated with the decision to place one's trust in a trustee.

Issues of trust, risk, and uncertainty thus appear to be intimately connected. Indeed, they might appear to be so intimately connected as to blur the distinction between trust and risk. Trust might thus be glossed as a form of risky decision-making, or decision-making under uncertainty. This would certainly appear to be consistent with our intuitions about the nature of trust dilemmas. As noted by Johnson-George and Swap (1982, p. 1306) a "willingness to take risks may be one of the few characteristics common to all trust situations." In this sense, it is natural to think of trustworthiness as a way of minimizing risk: the more trustworthy someone is, the more likely they are to behave in the way we want them to behave, and thus the less likely things are to go awry when the trustee has an opportunity to fulfil (or betray) the trust that we place in them.

Undoubtedly, there is something important about the relationship between trust, risk, and uncertainty, and it may be the case that risk and uncertainty are common to all trust situations/dilemmas. Some care is, however, required when it comes to understanding the nature of the relationship between trust, risk, and uncertainty. Consider, for example, that we have interpreted trust from a doxastic perspective; that is to say, we are suggesting that trust is a form of belief, specifically a belief about a trustee's trustworthiness. This commitment to trust as a belief constrains the way we think about the relationship between trust and risk. Consider, for example, that there is nothing inherently risky about my belief that you are either trustworthy or untrustworthy. Perhaps you are genuinely untrustworthy (relative to me), but there is no risk associated with my believing (falsely) that you are trustworthy. I can falsely believe that you are trustworthy when in fact you are not, but this does not mean that I am exposed to any risk in forming this belief about you. The notion of risk only comes into play

when I proceed to make a decision about whether or not to act on my trust and place my trust in you. At this point, I am undoubtedly involved in a form of risky decision-making, for if I make the wrong decision, then it could be very costly for me.

The result is that it only makes sense to talk about risk when a trust decision is made. We therefore need to make a distinction between trust beliefs (the beliefs that X has about Y 's trustworthiness) and trust decisions (the decisions that X makes on the basis of their beliefs about Y 's trustworthiness). This distinction is important, for we can clearly believe that someone (or something) is trustworthy without necessarily having to rely on them. I may implicitly trust Tesla self-driving cars, but if I never have an opportunity to use a Tesla car (perhaps because I know I will never be able to afford one), then I will never have an opportunity to put this trust to the 'test', so to speak.

What about uncertainty? Like risk, uncertainty appears to be a common feature of trust situations. It is, indeed, this uncertainty that underlies the risk that accompanies trust decisions. If there is no uncertainty about how things are to unfold, then we already know how things will unfold, and thus there is no risk in deciding whether or not to make ourselves vulnerable to a trustee's actions. In such situations, there seems little reason to invoke the notion of trust as a means of explaining/justifying the trustor's thoughts and actions towards the trustee. If I (as trustor) already know exactly how you (as trustee) will behave, and I also know that nothing will interfere with your behaviour, then I already know what the future will bring, and the risk is therefore nullified. In such situations, it is hard to see why I would say that I trust you. I place my trust in you because I know exactly how you *will* behave; I do not have to worry how you *might* behave. If someone asks me why I made myself vulnerable to you, I can simply refute the claim that I am vulnerable by referring to the fact that I already know how you will behave. There is no need, in this situation, for me to explain or justify my actions by appealing to the notion of trust. What legitimates the explanatory appeal to trust in our exchanges and dealings with one another is that human individuals are unruly beasts whose behaviour is seldom predictable in a way that transforms our *beliefs* about the future into *knowledge* of the future. As noted by Baier (1985, p. 61), "given the partial opaqueness to us of the reasoning and motivation of those we trust and with whom we cooperate," there is always a degree of risk involved in the placement of trust. We are never completely sure how another individual might behave, and this is so regardless of how trustworthy they might be.

It is important to note that there are a number of forms of uncertainty at work in trust situations. The most salient form of uncertainty relates to our uncertainty about the abilities, intentions, and motivations of the trustee. This is our basic level of uncertainty about the trustworthiness of the trustee. Call it trustee uncertainty. In situations where the trustee is unknown to the

trustor, then issues of trustee uncertainty are of paramount significance. As Gambetta (1988) aptly observed in an early and influential treatment of this problem:

The condition of ignorance or uncertainty about other people's behavior is central to thenotion of trust. It is related to the limits of our capacity ever to achieve a full knowledge of others, their motives, and their responses to endogenous as well as exogenous changes. (Gambetta, 1988, p. 218)

Of course, if we had enough information about the trustee, then we might be able to come to some more or less accurate judgement about their trustworthiness. Even here, however, things are not straightforward. For one's certainty about a trustee's trustworthiness is only as good as the information that informs the evaluative process. What if the information is inaccurate, out of date, or (perhaps worse) deliberately engineered so as to bias the outcome of the trustor's evaluative process?

In addition to trustee uncertainty, there is a further form of uncertainty that influences trust-related decisions (although not necessarily trust beliefs). This is what we will call situational uncertainty. Situational uncertainty refers to the trustor's uncertainty about how the present situation will evolve in the future, at least until the point where the trustee has an opportunity to fulfil (or betray) the trust that is placed in them. The problem, here, is that the situation may change in such a way as to undermine the assumptions that were made at the point trust beliefs were formed and trust decisions made.

There are two aspects to this situational uncertainty. The first is that the situation changes in such a way as to invalidate X's beliefs about Y's trustworthiness. If the incentive structure of the environment changes to the extent that Y is motivated to engage in a course of action that is counter to X's interests, then Y's trustworthiness has been affected by a shift in the situation. Situation-relevant changes thus threaten to undermine the integrity of X's trust-related cognitions, even if these cognitions were, at the time, perfectly accurate.

Another aspect to situational uncertainty relates to changes that cause accidental violations of trust. In this case, the perceived trustworthiness of Y may be unaffected; nevertheless, Y's capacity to φ is negated by some sort of change in the situation in which this capacity was supposed to be exercised. This shift in the situation may be due to forces and factors that lie beyond the control of both the trustor and the trustee, and, in this sense, they do not necessarily affect the trustee's trustworthiness. Nevertheless, X trusted Y to φ and φ failed to occur. The upshot is that X bears the cost of Y not φ-ing, even though Y is not to blame for the fact that φ failed to occur.

A consideration of situational uncertainty highlights the complex relationships between trust, uncertainty, and risk. Consider, for example, that trustee uncertainty may be minimal (perhaps

even zero), in which case, X can be said to possess complete certainty about Y's trustworthiness. In this situation, we might expect the risk to X to be minimal, and perhaps non-existent. Accordingly, we might expect them to act on their trust and place their trust in Y, for they already know that Y will behave in a manner that is consistent with their expectations. As should be clear from the discussion of situational uncertainty, however, even in situations where there is no trustee uncertainty, X is still required to make a risky decision simply because there are no guarantees that φ will actually occur. A failure to φ could occur for all manner of reasons, and only some of those are attributable to Y's trustworthiness (or lack thereof).

This insight is important, for it reminds us that even in situations where X trusts Y and does so with 100% confidence, there is no guarantee that X will proceed to place their trust in Y. If the situational uncertainty is sufficiently high, and the risks are sufficiently large, then X may refrain from placing their trust in Y simply because the risks are too high given the prevailing level of situational uncertainty. This risk has nothing to do with trustworthiness per se; it is more due to the inherent unpredictability of future states-of-affairs, which may be accentuated in certain kinds of social contexts (e.g., rapid social change). No amount of trustworthiness will necessarily reduce this risk, since the risk is not, in fact, attributable to trustworthiness. Instead, the risk is attributable to the situation, and the proper focus of attention is thus the stability and predictability of the environment in which both the trustor and trustee are embedded. This is not to say that trustee uncertainty and situational uncertainty are entirely distinct from one another. If we hold situational uncertainty constant, then it is easy to see why perceived trustworthiness would reduce the overall level of (perceived) risk associated with a trust decision. If situational uncertainty is zero, then all of the risk relates to the actions of the trustee: either Y does what they are trusted to do, or they do not. In this situation, the main risk for the trustor relates to errors in the assessment of a trustee's trustworthiness. If X trusts Y, and Y is genuinely (objectively) trustworthy, then (assuming the absence of situational uncertainty), there are no risks to X's decision to place their trust in Y. If X is uncertain about Y's trustworthiness, then the perceived risk of a trust-related decision increases with increasing uncertainty about the trust (again, assuming the absence of situational uncertainty). The addition of situational uncertainty complicates both the trust assessment and risk assessment processes. For a start, situational uncertainty elevates the risk associated with trust decisions, even if the trustee should be completely trustworthy. As we have seen, one reason for this is that it alters the forces and factors that underwrite Y's trustworthiness, which requires X to anticipate how Y's trustworthiness might change in the future. Secondly, even if Y's trustworthiness should remain unchanged, there is no guarantee that Y will be able to fulfil the trust that is placed in them.

Hopefully, this discussion sheds some light on the relationship between trust, risk, and uncertainty, and the way in which we might go about modelling such relationships. The notion of risk depends on the cost/benefit structure of a trust situation. Sometimes the stakes are high; sometimes they are not. If the stakes are sufficiently low, such that we are indifferent to the outcome of a trust decision (i.e., our decision to place to trust), then it is hard to see why we ought to appeal to the notion of trust in explaining our actions; for the explanatory appeal to trust only makes sense when there is some sort of cost associated with the failure to fulfil trust—where the actions (or inaction) of a trustee actually matters to us. This risk is no doubt tied to uncertainty, in the sense that the less certain we are about how things will pan out, the more risk is involved in placing trust. This risk is, however, informed by different forms of uncertainty: the uncertainty we have about the actual trustworthiness of a trustee and the uncertainty we have about the specific features of the situation in which the trustee will attempt to fulfil the trust we place in them. Perhaps unsurprisingly, the literature on trust has focussed on one particular kind of uncertainty, namely, the uncertainty that stems from a consideration of another's trustworthiness (i.e., trustee uncertainty). This is, to be sure, an important form of uncertainty, and it is undoubtedly one that influences trust-related decisions and actions. But it is not the only kind of uncertainty that affects such decisions and actions, nor is it necessarily the most important form of uncertainty when it comes to understanding the forces and factors that enable productive forms of social cooperation to flourish.

6 Trust Process Model

We have suggested that trust ought to be conceptualized as the beliefs that a given trustor has about a given trustee's trustworthiness relative to particular activities and the circumstances in which those activities are to be performed. These beliefs are what we have called *trust beliefs*. Trust beliefs are to be contrasted with *trust decisions*, which relate to the decision about whether or not to act on one's beliefs about the trustee's trustworthiness.

From a process-level perspective, this characterization suggests that there are three steps associated with the act of placing trust. The first of these steps is the process that leads to the formation of trust beliefs. This belief-forming process is what we call trust assessment. The goal of trust assessment is to assess or evaluate the trustworthiness of a given trust object or trustee. The output of this process is a belief about the trustee's trustworthiness, i.e., a trust belief.

It is important to note that trust beliefs are distinct from trust actions, which refers to the actions associated with the placing of trust. As noted above, one could possess beliefs about the

trustworthiness of a trustee without ever having an opportunity to place their trust in the trustee. In this case, there would be no risk to the trustor, for the notion of risk only comes into play when the trustor places their trust in the trustee and thereby makes themselves vulnerable to the trustee.[4]

Trust beliefs are, we suggest, the input for a decision-making process whose output is the decision about whether or not to implement trust actions. If a trustee believes that a trustee is trustworthy and decides to act on their trust, then they will, in all likelihood, attempt to act on their trust by implementing trust actions.

To help us understand this three-stage process, consider a situation in which a NIS stakeholder is confronted with a candidate NIS. Among other things, the stakeholder (X) will want to gauge the trustworthiness of the NIS (Y). In order to do so, they will engage in the process of trust assessment. The output of this process will be a trust belief that expresses the level of trust X has in Y's trustworthiness. X may then proceed to decide whether or not they want to commit themselves to the use of the NIS. X's trust beliefs will feed into this process and culminate in a decision about whether to commit to the use of the NIS. Finally, the act of committing to the use of the NIS (e.g., purchasing or deploying the NIS) corresponds to the trust action.

One of the hazards of this sort of characterization is that it appears to suggest that the three-step process can be subdivided into a cognitive component (trust assessment and trust decision-making) and an action component (the implementation of trust actions). Relative to recent work in cognitive science, however, it is by no means clear that the putative distinction between cognition and action is as straightforward as it might appear (see Clark, 2008). An additional concern is that it is by no means clear what is entailed by the notion of trust assessment. What sort of process is this, exactly? It is no doubt tempting to regard trust assessment as a complex form of knowledge- guided reasoning process—one that incorporates available information for the purposes of deriving a rational estimate of a trustee's trustworthiness. On the other hand, trust assessment may be more akin to a perceptual process, in the sense that a trustee's trustworthiness is simply perceived based on whatever information is available to hand. As regards the cognition/action distinction, we suggest that the process of trust assessment (in its various guises as either a perceptual or reasoning process) may be implemented in either an active or passive manner. What we mean by

[4] As noted by PytlikZillig and Kimbrough (2016), it is unclear whether trust beliefs can be treated independently of risk. This is because, in evaluating the trustworthiness of a trustee, the trustor is obliged to attend to the hazards associated with the negative outcomes of ill-placed trust (see also Cao, 2015). It could thus be argued that risk is not irrelevant to the formation of trust beliefs, for while there is nothing inherently risky about the mere belief that someone is trustworthy or untrustworthy, assessing someone's trustworthiness does require one to imagine the consequences of placing trust in the wrong trustees.

passive, in this context, is that a trustor performs a trust assessment process with whatever information is available to them at a particular time; they make no effort to solicit more information from the external environment over and above what is already available to them at the time the trust assessment process is performed. This passive form of trust assessment contrasts with what we call the active form of trust assessment. In this case, the trustor plays an active role in soliciting information from the environment so as to meliorate the trust assessment process. The trustor might, for example, seek additional information about the trustee, or they may probe and prod the trustee with particular questions. They may even seek to evaluate trust-related hypotheses by testing the trustee in various ways. In all the cases, the trustor is an active player in the process by which assessments of trustworthiness are made. They engage in a variety of what cognitive scientists call epistemic actions (see Kirsh & Maglio, 1994), each of which is intended to yield access to certain bodies of trust-relevant information.

How ought we to accommodate the notions of risk and uncertainty into this trust process model? Given the exploration of the relationship between trust, risk, and uncertainty in Section 6, it seems reasonable to mark a distinction between the process of trust assessment, on the one hand, and the process of risk assessment, on the other.

Focusing first on the process of trust assessment, we assume that a trustor will process information that is relevant to the determination of the trustee's trustworthiness. Some of the criteria that are likely to be important in the context of NISs include information pertaining to security, privacy, ethicality, robustness, resilience, reliability, legality, verifiability, functionality, responsibility, interoperability, and so on. There are a number of challenges associated with the processing of this information. Firstly, the trustor needs to assess the reliability of the information. Information about the security of a NIS may be scored very highly in terms of its completeness and detail, but unless the information is deemed to be reliable, then it will count for naught. This is a particular problem in trust assessment scenarios. The primary aim of the trustor is to evaluate the trustworthiness of a trustee by minimizing trustee uncertainty. Following Bacharach and Gambetta (2001), this is what we might call the primary problem of trust. Unfortunately, as noted by Bacharach and Gambetta (2001), the primary problem of trust is not the only trust-related problem hereabouts, nor is it necessarily the most challenging problem confronting the trust assessment process. The problem is that the reliability of the trust assessment process (i.e., the ability of the trust assessment process to deliver an accurate judgement about the trustee's trustworthiness) is bound up with the reliability of the information that informs the trust assessment process. This is what Bacharach and Gambetta (2001) refer to as the secondary problem of trust. The secondary problem of trust is, in essence, the problem of evaluating the reliability of the information that ultimately fixes the

content of trust beliefs. This problem is particularly awkward in trust-related situations, for trustees may have an incentive to provide inaccurate information for the purposes of influencing the outcome of the trust assessment process.

Assuming that information about the trustee is reliable, then the primary constraints on the trust assessment process are the judicious selection of appropriate information, the availability of such information, and the appropriate integration of this information into some rationale, reason- respecting chain of inference. In many cases where the reliability of information is not a concern, the availability of information is likely to be a predominant concern. These include cases where information is not presented with sufficient detail to support or trust assessment process, or where information is unavailable due to restricted access (e.g., concerns about cyber-security).

The primary risk associated with the trust assessment process is a failure to accurately assess the trustworthiness of a trustee. As noted above, trust beliefs are not, by themselves, inherently risky: One can believe that a trustee is trustworthy, even though they are not. But providing these beliefs do not get translated into trust actions, then the trustor is not put at risk as a result of their beliefs.

There is, however, a sense in which trust beliefs are risky. If a trustor misjudges a trustee's trustworthiness and this leads them to place trust in an untrustworthy trustee, then the costs to the trustor could be considerable. The accurate assessment of trustworthiness is thus a means of reducing risk in situations where one is inclined to commit to a trust action and thus make themselves vulnerable to a trustee. As noted above, however, it is hard to see why these risk-related concerns would be a feature of trust assessment. The goal of trust assessment, recall, is to assess the trustworthiness of a trustee. Arguably, this is not a process that ought to be overly concerned with the risks associated with a subsequent decision. The notion of risk comes into play once we proceed from the process of trust assessment to the process of trust decision-making. At this point, the trustor is required to assess the risks associated with various courses of action. Such risks are, no doubt, informed by the outputs of the trust assessment process, but they are not limited to such outputs; nor is there any reason to assume that the outputs of the trust assessment process are the most important determinant of whether or not a trust-related decision (yea or nay) is made.

7 Trust Assessment

As we have seen, trust assessment is the process that culminates in the formation of a belief about a trustee's trustworthiness. In an interpersonal context, this process plays a crucial

cognitive role in enabling us to discriminate between the trustworthy from the untrustworthy—those individuals that we can rely on to further our own interests and those who should probably be avoided at all costs.

Within the scientific and engineering community, there has been a concerted effort to understand the forces and factors that shape the process of trust assessment, and one of the goals of computationally-oriented work in this area has been to formalize the process of trust assessment with a view to subjecting trustees to automated forms of trust evaluation (Truong, Lee, Askwith, & Lee, 2017). One of the challenges confronting work in this area is that the information required to support the trust assessment process often varies from one context to another, requiring analytic and synthetic efforts to be restricted to specific task and application contexts. At a general level, however, there has been considerable progress in understanding the sorts of information that guide trust-related evaluations. In this section, we discuss some of the more prominent strands of research that speak to this issue.

7.1 Trustworthiness Attributes

The trust literature uses a variety of terms to refer to the forces and factors that influence the trust assessment process. In the context of the present report, we make a distinction between trustworthiness attributes, trust indicators, and trust-warranting properties. Trustworthiness attributes are the attributes of a trustee that speak to the trustee's trustworthiness.

Reviewing the results of work in the social sciences, Cook and Gerbasi (2009) suggest that there are at least two factors that appear to be relevant to assessments of trustworthiness. These are: 1) the competence or reliability of the trustee (Can the trustee be expected to φ simply on the basis of their ability?), and 2) the integrity, honesty, and benevolence of the trustee (Can the trustee be expected to 'do not any harm', at least with respect to trustor?).

A somewhat more refined approach to trustworthiness attributes is provided by Mayer et al. (1995). Mayer et al. (1995) are primarily concerned with the forces and factors that influence judgements of trustworthiness in organizational settings, such as in the workplace; nevertheless, their analysis is one that has been applied to many trust-related situations, including those in which the trustee is a non-human entity, such as technological system.

Mayer et al. (1995) offer a three-factor approach to trustworthiness, which has come to be known as the ABI model of trustworthiness. According, to this model, trustworthiness is a function of the following properties:

- **Ability/Competence/Capacity:** In order for X to trust Y, X must believe that Y is

capable of φ. If you know that I am capable of doing something, then you are more likely to trust me.

- **Benevolence/Care/Concern:** Expressions of care, concern, and prosociality typically enhance perceptions of trustworthiness. If you know that I care about you, then you are more likely to trust me.
- **Integrity/Honesty/Sincerity:** X's beliefs about Y are based on information about Y. If X discovers that Y has deliberately communicated false information, then X will not trust Y. You cannot trust me if I lie to you about my abilities. Similarly, you cannot trust me if my expressions of concern towards you are revealed to be insincere.

Table 1. Some of the terms used to refer to the ability, benevolence, and integrity aspects of the Ability, Benevolence, and Integrity (ABI) model (source: Truong et al., 2017).)

Ability	Benevolence	Integrity
Competence, ability, capability, expertness, credibility, predictability, timeliness, robustness, safety, stability, scalability, reliability, dependability.	Good intention, goodness, certainty, cooperation, cooperativeness, loyalty, openness, caring, receptivity, assurance.	Honesty, morality, completeness, consistency, accuracy, certainty, availability, responsiveness, faith, discreetness, fairness, promise fulfilment, persistence, responsibility, tactfulness, sincerity, value congeniality, accessibility.

One of the problems with the ABI model relates to the difficulties in distinguishing between the ability, benevolence, and integrity dimensions. Consider, for example, some of the terms that are frequently associated with each of these dimensions (see Table 1). Some of these terms are relatively easy to associate with one of the dimensions; others are not. It remains unclear whether the terms "ability," "benevolence," and "integrity" are the best linguistic labels for dimensions of trustworthiness; nevertheless, Mayer et al. (1995) suggest that, when taken together, these three factors appear to explain a significant proportion of the variability when it comes to judgements of trustworthiness.[5]

The ABI model is readily applicable to paradigmatic trust relationships centred on individual human agents. It is, however, much harder to apply the ABI model to situations involving non-human trustees, such as institutions, technological systems, artificial agents, and so on. One

[5] In particular, Mayer et al. (1995, p. 722) suggest that, "Each element contributes a unique perceptual perspective from which the trustor considers the trustee. If a trustee is perceived as high on all three factors, it is argued here that the trustee will be perceived as quite trustworthy."

reason for this is that it is hard to see how such systems could be seen to possess any form of benevolence or integrity, at least as these terms are applied to human individuals. Relative to NISs, the most prominent target of the trust assessment process is likely to be directed at the ability dimension (i.e., trust assessment is likely to be concerned with issues of ability, capability, and competence). Is this a problem for the idea that NISs ought to be seen as trustworthy? Does, for example, the putative absence of benevolence and integrity impugn the trustworthy status of a trustee? One way of answering this question is to assess the extent to which issues of benevolence and integrity are a necessary ingredient of human trust relationships. While a great many human trust relationships might be predicated on the trustee's benevolence and integrity, it is not clear that *all* such relationships are based on such properties. In paying for professional services, for example, there are surely situations in which we choose one service provider over another based simply on the basis of ability-based considerations without regard to how those providers might feel towards us.

In any case, it is by no means clear that attributions of benevolence and integrity are entirely without merit in trust relationships involving non-human counterparts. We might judge a system to be benevolent in the sense that it is a benign system that poses no real threat to us as individuals. Inasmuch as a NIS is designed to operate in a manner that does no harm (and perhaps even contributes to the common good), then it might be regarded as benevolent in precisely this sort of sense.

7.2 Trust Indicators

Trustworthiness attributes refer to the properties of a trustee that are relevant to the trustee's trustworthiness. For the most part, however, our access to these attributes is somewhat indirect. In assessing the trustworthiness of a surgeon, for example, we generally do not demand evidence of the surgeon's ability to perform a given surgical procedure. Instead, we rely on information that indicates the presence of this ability. If the surgeon is a qualified surgeon, then it is natural to assume that they must possess a basic level of surgical competence. This makes sense, for if they did not possess this competence, then it is hard to see how they would have secured the qualifications that enabled them to practice surgery. This highlights a distinction between what we are calling trustworthiness attributes, which are properties of the trustee, and trust indicators, which indicate the possession of these properties.

7.3 Trust Warranting Properties

In addition to trustworthiness attributes and trust indicators, the trust literature frequently makes an appeal to so-called trust-warranting properties. Trust-warranting properties are

typically understood as the forces and factors that influence the trustee in such a way as to encourage trust- fulfilment. An excellent overview of some of the more prominent examples of trust-warranting properties is provided by Riegelsberger et al. (2005). They divide trust-warranting properties into two categories, namely, contextual properties and intrinsic properties, where intrinsic properties are what we have dubbed trustworthiness attributes and/or trust indicators. The more interesting category of trust-warranting properties, at least for present purposes, are contextual properties, which are glossed as "factors that can induce such an actor to behave in a trustworthy manner" (Riegelsberger et al., 2005, p. 393). The following are the contextual properties identified by Riegelsberger et al. (2005):

- **Temporal Embeddedness:** Temporal embeddedness refers to the extent to which a trust relationship extends into the future. If X has an opportunity to trust Y on multiple occasions, then X has an opportunity to learn from their experience of trusting Y. Perhaps more importantly, the possibility of future cooperation serves as an incentive for Y to fulfil the trust that has been placed in them. If Y knows that they will benefit from being trusted by X, then they have an incentive for living up to X's expectations of them.
- **Social Embeddedness:** Social embeddedness refers to the extent to which X and Y are embedded in social networks that support the propagation of information about Y 's behaviour. If Y fails to fulfil the trust that is placed in them, then X has an opportunity to communicate this information to other would-be trustors, thereby inflicting reputational damage on Y.
- **Institutional Embeddedness:** Institutional embeddedness refers to the presence of institutions that impose constraints on the behaviour of individuals and organizations. Examples include the likes of law enforcement agencies, judicial systems, trade organizations, and so on. Institutions can sometimes act as trust brokers that certify the trustworthiness of a trustee in the absence of any information that a trustor may have about a trustee.

All these forms of 'embeddedness' are potentially applicable to NISs, although the notion of institutional embeddedness is perhaps the more important form of embeddedness, in the sense that NISs might be expected to conform to (e.g.) legal constraints.

8 National Identity Systems: Trustees

One of the first issues to be addressed as part of attempts to model the trust assessment process is: “Who is the trustee?” In the context of NIS, this question might seem to have a relatively straightforward answer. The answer, presumably, is the NIS.

In fact, the nature of the trustee in many technological contexts (and NISs are no exception) is rarely straightforward. One immediate problem concerns the nature of the system itself. Is the system to be regarded as a purely technological system, or is it better cast as a socio-technical system? Inasmuch as it is a socio-technical system, then its functionality may depend, at least in part, on the activities of multiple human individuals. In this case, it is not immediately clear how we ought to assess the trustworthiness of a system. Should we, for example, decompose the larger systemic organization into technological and social parts and subject these parts to separate trust- related evaluations? Or should we aim to direct the assessment process towards the larger, hybrid organization? Even if the system to be evaluated is a purely technological system, one whose functionality does not supervene on the activities of multiple human individuals, then we are still left with the basic problem of who (or what) ought to be subject to trust-related evaluative scrutiny. In respect of NIS, for example, we could restrict the focus of our attention to the technological system, but there are surely grounds for thinking that the individuals and organizations responsible for the development of a NIS are also contributing to the trustworthiness of the system. After all, the functionality of the NIS was produced as a consequence of the activities of these individual and collective agents, so in as much as we trust these agents, then perhaps we ought also to trust the technological artefacts that those agents produce. Other potential candidates for trust evaluation include those responsible for maintaining and updating the systems, those who manage and control the system, and those who might be in a position to exploit the system for socially beneficial or socially malignant purposes. To the extent that we want to base trust decisions on the behaviour of a NIS, as well as its consequences for individuals, groups, and societies, then there is surely a *prima facie* case for considering these individuals as *bona fide* targets of the trust assessment process.

9 National Identity Systems: Stakeholder Perspectives

In addition to asking questions about who or what is the trustee in the context of NISs, it is also possible to ask questions about the nature of those with an interest in assessing the trustworthiness of NISs. As with the trustee-related question, this question has a seemingly straightforward answer. The trustors are those agencies who are required to make decisions about the adoption and deployment of NISs.

Again, however, this response overlooks the potential for NISs to be evaluated from different perspectives and by different stakeholder groups. Relative to the earlier discussion on selective commitments and target audiences (see Section 4), it should be clear that NISs (and other technological systems) need not be equally trustworthy to the members of different communities. Much will depend on the extent to which the functionality of the system is compatible with the interests of those who might be affected by the introduction of such systems. This issue is particularly important when it comes to NISs, for such systems are apt to exert widespread influences across national populations, and, once deployed, it may be difficult to impossible to opt- out of such systems.

A multi-stakeholder approach to the trust-related evaluation of NISs raises a number of issues and concerns, some of which are relevant to the modelling of trust assessment processes. Given the relativistic nature of trustworthiness (see Section 2), it is likely that different stakeholder groups will evaluate identical systems in different ways, either by drawing on different bodies of trust- relevant information or assigning greater weight to some features over others. This calls for a degree of flexibility with regard to the formal modelling of the trust assessment process. One way of modelling trust assessment from a knowledge-based perspective is via the specification and selection of 'norms', where a norm is equivalent to the notion of an evaluative criterion (see Schreiber et al., 2000, pp. 134–136). As noted by Schreiber et al. (2000), the process of norm selection can sometimes be cast as a form of knowledge-intensive process, one which draws on background knowledge to select and prioritize norms based on the features of the assessment process or the broader context in which the assessment process occurs.

Another issue to consider relates to individual differences between trustors. One way of understanding these differences is via the notion of trust propensity, which is deemed to influence the extent to which an individual trustor is disposed to regard a trustee as trustworthy in the absence of any specification information about that trustee. Mayer et al. (1995) incorporate the notion of trust propensity into their ABI model of trust/trustworthiness (see Section 8.1). According to Mayer et al. (1995): Propensity to trust is proposed to be a stable within-party factor that will affect the likelihood the party will trust. People differ in their inherent propensity to trust.

Propensity might be thought of as the general willingness to trust others. Propensity will influence how much trust one has for a trustee prior to data on that particular party being available. People with different developmental experiences, personality types, and cultural backgrounds vary in their propensity to trust (e.g., Hofstede, 1980). (Mayer et al., 1995, p. 1995)

This appeal to cultural backgrounds has a particular significance in the context of NISs, for it raises an important question about the extent to which trust assessment ought to be regarded as a culturally neutral phenomenon, something that is unaffected by the cultural idiosyncrasies of disparate nation states. If this should turn out *not* to be the case—if, for example, the process of trust assessment varies in a systematic fashion across culturally-circumscribed communities—then it is unclear that a single one-size-fits-all type approach to the development of dissemination of trustworthy NISs can be made to work.

There are two additional issues to consider, here, both of which are tied to the notion of culturally- idiosyncratic modes of trust evaluation. The first relates to nation states whose citizens are drawn from multiple cultural, ethnic, linguistic, and religious communities. It is, at present, unclear how these sorts of differences might affect the trust-related evaluation of technological systems that are designed to apply, in a rather uniform fashion, to all the citizens of a nation state. An additional complicating factor relates to the foundational notions of trust and trustworthiness. Much of the work relating to trust and trustworthiness has been undertaken from a broadly Western perspective, and it is easy to assume that trust-related concepts are simply part of the common conceptual furniture of the hominin social world—part of the conceptual backdrop against which our species-specific social interactions and exchanges occur. This may be something of an optimistic assumption, however. As noted by Hardin (2002), there are variations in the use of trust-related terms across cultures, and some languages (e.g., French) do not have a specific word for trust. This raises an issue about the cultural neutrality of attempts to subject the notions of trust and trustworthiness to analytic scrutiny. Inasmuch as these analytic efforts yield a culturally- inflected, and thus culturally-biased, portrayal of what trust is, then the attempt to provide an all- encompassing theoretical account of trust—one that is then foisted upon foreign nation states(!)— amounts to little more than a form of conceptual colonialism. Given the international orientation of the Trustworthy Digital Infrastructure for Identity Systems initiative, this is undoubtedly an important area for future research and investigation.

10 Conclusion

In the present report, we outlined a doxastic approach to trust that situates trust-related concepts within the folk psychological explanatory frameworks that are used to make the behaviour of ourselves and others intelligible to both ourselves and others. According to this approach, trust is to be conceptualized as a belief about the trustworthiness of a trustee. Trustworthiness, by contrast, is a property of a trustee (a trustee being the object of trust and the target of trust evaluation).

In addition to discussing some of the more general issues raised by the analysis of trust-related concepts (see Section 2), we also sought to outline a parametric approach to the modelling of trust- related concepts. This approach is intended to obviate some of the definitional difficulties that have plagued previous analytic efforts (see Section 4 and Section 5). Additional contributions of the present report include an exploration of the relationship between trust, uncertainty, and risk (see Section 6), a process-level model of trust-related cognitions (see Section 7), an analysis of the factors that are likely to be relevant to the trust-related evaluation of NISs (see Section 8), and a consideration of some of the issues raised by a multi-stakeholder approach to trust evaluation (see Section 10).

The present report summarizes work that has been undertaken in respect of the RM-NIS project component of the Trustworthy Digital Infrastructure for Identity Systems initiative, which is led by the Alan Turing Institute. As part of our future work on the RM-NIS project, we aim to extend and formalize the conceptual approach described in the present report so as to support the trust-related evaluation of NISs from a multi-stakeholder perspective.

References

Bacharach, M., & Gambetta, D. (2001). Trust in Signs. In K. S. Cook (Ed.), *Trust in Society* (pp. 148–184). New York, New York, USA: Russell Sage Foundation.

Baier, A. (1986). Trust and antitrust. *Ethics*, 96 (2), 231–260.

Baier, A. C. (1985). What do women want in a moral theory? *Noûs*, 19 (1), 53–63.

Cao, L. (2015). Differentiating confidence in the police, trust in the police, and satisfaction with the police. *Policing: An International Journal of Police Strategies & Management*, 38 (2), 239–249.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York, New York, USA: Oxford University Press.

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, 36 (3), 181–204.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action and the Embodied Mind*. New York, New York, USA: Oxford University Press.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, Massachusetts, USA: Harvard University Press.

Cook, K. S., & Gerbasi, A. (2009). Trust. In P. Hedström & P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology* (pp. 218–241). Oxford, UK: Oxford University Press.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13 (7), 293–301.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 127–138.

Gambetta, D. (1988). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relationships* (pp. 213–237). Cambridge, Massachusetts, USA: Basil Blackwell.

Giddens, A. (1990). *The Consequences of Modernity*. Stanford, California, USA: Stanford University Press.

Goldberg, S. C. (2020). Trust and Reliance. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 97–108). New York, New York, USA: Routledge.

Grodzinsky, F., Miller, K., & Wolf, M. J. (2020). Trust in Artificial Agents. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 298–312). New York, New York, USA: Routledge.

Hardin, R. (2001). Conceptions and Explanations of Trust. In K. S. Cook (Ed.), *Trust in Society* (pp. 3–39). New York, New York, USA: Russell Sage Foundation.

Hardin, R. (2002). *Trust and Trustworthiness*. New York, New York, USA: Russell Sage Foundation.

Hardin, R. (2013). Government without trust. *Journal of Trust Research*, 3 (1), 32–52.

Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, 48 (1), 1–20.

Hinton, G. E. (2010). Learning to represent visual input. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365 (1537), 177–184.

Hofstede, G. (1980). Motivation, leadership, and organization: Do American theories apply abroad? *Organizational Dynamics*, 9 (1), 42–63.

Hohwy, J. (2013). *The Predictive Mind*. Oxford, UK: Oxford University Press.

Johnson-George, C., & Swap, W. C. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, 43 (6), 1306–1317.

Jones, K. (1996). Trust as an affective attitude. *Ethics*, 107 (1), 4–25.

Keren, A. (2014). Trust and belief: a preemptive reasons account. *Synthese*, 191 (12), 2593–2615.

Keren, A. (2020). Trust, Preemption, and Knowledge. In K. Dormandy (Ed.), *Trust in Epistemology* (pp. 114–135). New York, New York, USA: Routledge.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513–549.

Lahno, B. (2020). Trust and Emotion. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 147–159). New York, New York, USA: Routledge.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20 (3), 709–734.

McLeod, C. (2020). Trust. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Stanford, California, USA: Stanford University.

O'Hara, K. (2012). *A General Definition of Trust* (Tech. Rep.). Southampton, UK: Electronics and Computer Science, University of Southampton.

PytlikZillig, L. M., & Kimbrough, C. D. (2016). Consensus on Conceptualizations and Definitions of Trust: Are We There Yet? In E. Shockley, T. M. S. Neal, L. M. PytlikZillig, & B. H. Bornstein (Eds.), *Interdisciplinary Perspectives on Trust: Towards Theoretical and Methodological Integration* (pp. 17–47). London, UK: Springer.

Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. *International Journal of Human–Computer Studies*, 62 (3), 381–422.

Robbins, B. G. (2016). What is trust? A multidisciplinary review, critique, and synthesis. *Sociology Compass*, 10 (10), 972–986.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N. R., Van de Velde, W., & Weilinga, B. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. Cambridge, Massachusetts, USA: MIT Press.

Smart, P. R. (in press). Predicting Me: The Route to Digital Immortality? In R. W. Clowes, K. Gärtner, & I. Hipólito (Eds.), *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts*. Berlin, Germany: Springer.

Smart, P. R., & Clowes, R. W. (2021). Intellectual Virtues and Internet-Extended Knowledge. *Social Epistemology Review and Reply Collective*, 10 (1), 7–21.

Truong, N. B., Lee, H., Askwith, B., & Lee, G. M. (2017). Toward a Trust Evaluation Mechanism in the Social Internet of Things. *Sensors*, 17 (6), 1346.