Digital Enlightenment Yearbook 2012
J. Bus et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-057-4-169

Trust from the Enlightenment to the Digital Enlightenment

Kieron O'HARA[1]
School of Electronics and Computer Science, University of Southampton

Abstract. A conceptual analysis of trust in terms of trustworthiness is set out, where trustworthiness is the property of an agent that she does what she claims she will do, and trust is an attitude taken by an agent to another, that the former believes that the latter is trustworthy. This analysis is then used to explore issues in the deployment of trustworthy digital systems online. The ideas of a series of philosophers from the Enlightenment – Hobbes, Burke, Rousseau, Hume, Smith and Kant – are examined in the light of this exploration to suggest how we might proceed in the Digital Enlightenment to ensure that systems are both trustworthy and trusted.

Keywords. Trust, trustworthiness

Introduction

Trust has been an enormous topic in the study of the World Wide Web, as evinced by the number of scholarly papers, the number of workshops associated with Web and Semantic Web conferences, and the prominence of trust in the layered model of the Semantic Web. Beyond that trust is a complex social phenomenon with a range of interpretations and understandings which have generally emerged by deriving theories of trust from paradigm cases, or particular social theories. As trust is deeply context-dependent, such paradigms have many idiosyncratic features, and the social setting often gets in the way of the important commonalities. The result is that much is often left out.

For example, on Niklas Luhmann's influential account trust is "an effective form of complexity reduction" [20], which is of course true, but is this a *defining* condition? There are many other reasons to trust. Many theories are built upon an analysis of face-to-face trusting relations between individual humans, which may or may not extrapolate to the World Wide Web [26]. Others take a 'revealed preference' attitude to trust (i.e. that taking part in an interaction entails trust in one's partners), which makes empirical investigation easier but creates a reductive equivalence between trust and the assumption of risk. We have to ask ourselves if this is a helpful way of looking at trust; for instance, if I give someone a 5* rating on eBay, I have thereby behaviourally demonstrated my trust in him or her, but I have not thereby taken a risk (although I presumably did take a risk earlier during our eBay transaction).

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

The lesson for Web Science is to avoid the trap of extrapolating *sui generis* theories of Web trust from particular areas of Web experience (such as social networking, or cybercrime). Much of human life is on the Web, and as new applications emerge and the Anglo-American-European bias of the Web declines, theories of Web trust will need sensitivity to many unpredictable and culturally-relative factors. The Web is not a technological realm in isolation; it is increasingly integrated into our work and leisure lives. Understanding trust on the Web involves understanding trust in human life.

A further error that is often made is to assume trust is a benefit, which should be unconditionally promoted: "how do we increase trust?" For instance, Fukuyama's influential account argues that "in all successful economic societies [economic] communities are united by trust" ([6], p.9), and that therefore we should aim for a high trust society. This is only partly true. Trust makes no sense, as I shall argue, without the prior concept of trustworthiness. The problem of trust, then, is emphatically *not* to increase trust unconditionally, but rather to link trust and trustworthiness effectively. Fukuyama should have spelt out a twofold aim: (i) become a high *trustworthiness* society, in which (ii) trust is placed effectively. Similarly, Luhmann's claim that trust reduces complexity can only be true when trust is well-placed (badly-placed trust will make things more difficult). This is as true online as anywhere else; the key requirement is to enable causal links between trust and trustworthiness.

In this paper I shall set out a conceptual analysis of trust (and trustworthiness) to underpin a range of trust relations, across not just human agents but non-human agents as well [24]. It is meant to apply to both rational and irrational decisions to trust. Its purpose is to highlight the key parameters that need to be accounted for when considering a trust relation, and which hopefully will feed into the design and development of Web technologies.

The concept of the Digital Enlightenment is a fascinating one, with the Web as a means of recreating the public space that allowed Enlightenment thought to flourish [6], [22]. Indeed, the Web fills the role allotted in the previous era to Diderot's *Encyclopédie*, "to collect all the knowledge scattered over the face of the earth, to present its general outlines and structure to the men with whom we live, and to transmit this to those who will come after us" ([5], p.17). Trust and trustworthiness were problematic for those Enlightenment *philosophes* who desired to place them on a rational foundation. Someone who places trust in someone else usually (not always) accepts vulnerability; it is hard, with a rationalist individualist moral psychology, to explain why anyone under those circumstances would take the trouble to be trustworthy. This raises a number of difficult philosophical issues of course [14], but is problematic for the Web as a socio-technical construct in particular, as any implementable models of interaction between people and machines must assume rationality as a principle ([31], pp.26-27). I shall discuss modern day online trust in the context of theories of trust and trustworthiness from the Enlightenment period, taking ideas from Thomas Hobbes (1588-1679), Edmund Burke (1729-97), Jean-Jacques Rousseau (1712-78), David Hume (1711-76), Adam Smith (1723-90) and Immanuel Kant (1724-1804) and recasting them for the Digital Enlightenment.

1. A Definition of Trust

In this section I will set out an analysis of trust in more detail. This will be in four parts, an analysis of trustworthiness, upon which will be built an analysis of trust. A discussion of failures of trust and trustworthiness follows, and finally I shall discuss issues surrounding the connection of the two. The analyses of trust and trustworthiness are developed in more detail in a working paper [24].

1.1. Trustworthiness

Trustworthiness is prior to trust, which is an attitude toward the trustworthiness of others. Indeed, as Hardin has argued [9], [10], many supposed commentators on trust are actually discussing trustworthiness. What, then, is this prior concept?

A trustworthy person is someone who does what she says she will do, all things being equal. This characterisation conceals quite a lot of structure. First of all, trustworthiness is a *property* of an *agent*. A *claim* must be made about her future actions. After all, it is absurd to accuse Barack Obama of being an untrustworthy brain surgeon, because he has never claimed to have brain surgery skills. The claim will also have the effect of narrowing the scope of trustworthiness; put another way, trustworthiness is context-dependent. The 'all things being equal' clause means that a trustworthy person need not succeed in carrying out the claimed behaviour, but if she does not, there must be an explanation for her failure which will absolve her of responsibility.

We can therefore define trustworthiness as a four-place relation, as follows:

> Y is trustworthy $=_{df}$ Tw<Y,Z,R,C> (1)
> where Y and Z are agents, R is a representation of the claim and C is a (task) context in which it applies

In (1), Y is the agent who, if (1) is true, is trustworthy. R is the content of the claim made about her[2] intentions, capacities and motivations in future behaviour; when (1) is true, Y's behaviour will be constrained by R. R may be explicitly written down, or may be implicit and understood; it may be open-ended and deliberately left unspecific to degrade gracefully. C is the set of contexts in which R is intended to apply (for instance, Y may claim to be a trustworthy car mechanic, but only within office hours, and only on certain makes of car).

This leaves Z, who is the agent responsible for generating and disseminating the claim R. In many, perhaps most, circumstances, Y = Z. However, this need not be the case. A trustworthy customer service employee respects a role description generated by her company. A trustworthy piece of software performs according to a specification written by a designer. It is essential that Z is *authorised* to make the claim about Y. Without authority, Z's claim has no bearing on Y's trustworthiness.

[2] I shall use these and other variables consistently throughout this paper to refer to occupiers of the various roles delineated. Additionally, to help defuse ambiguities I shall use masculine pronouns for the trustor (who will be christened X) and feminine pronouns for the trustee Y. To underline an obvious point to those who do not care to distinguish between grammatical and sexual gender, I do not thereby intend to suggest that only men trust and that only women are trustworthy.

1.2. Trust

Given that trustworthiness is a property of an agent, this leads us to the question of how that relates to trust. This is straightforward: trust is an *attitude* toward the trustworthiness of another. In other words, X trusts Y iff he believes that she is trustworthy.[3]

This is still a complex idea, however. Not only does trustworthiness bring with it context-dependency, but trust forces us to confront a subjective element. Trust is not as simple as X believing (1). Broadly speaking, there are six parameters of consequence in the trust relation, as follows:

$$\text{X trusts Y} =_{df} \text{Tr}\langle X,Y,Z,I(R,c),Deg,Warr\rangle \quad (2)$$
with Y, Z and R as before, and X an agent

In (2), the first three parameters are the relevant agents. X is the trustor, and Y the trustee. Z, as before, is the agent who makes the claim R about Y's intentions, capacities and motivations. And again, as before, it could be that Z = Y (it could also be that X = Y and X = Z as well, although we don't need the details of these identities here [24]).

Z makes a claim that Y's behaviour, all things being equal, will conform to R in contexts C. X's trust, if well-placed, should accept that claim. However, it need not, because X is only boundedly rational and communications between Z and X could fail. Furthermore, as noted above, R might be implicit or unspecific. Hence X has to *interpret* R for the contexts in which he is interested. I have written this as a function I(R,c), to be read as an interpretation of the force of R in the set of contexts that interest X, which I term c.

This brings trust's subjective aspect to the fore – for X's trust, it is *X's interpretation* that counts, whether or not it is correct. That highlights further restrictions on trust. As it is an attitude held by X about Y, it is X who supplies the underlying assumptions of the judgment. This has two specific consequences. First, for X to trust Y, it need not be the case that Z *has* authority to make claim R about Y. It is necessary and sufficient that X *believes that* Z has that authority. Second, I(R,c) only has any force with respect to Y if $c \subseteq C$, otherwise it will fall out of the scope of R. Yet for X's trust, it is necessary and sufficient *only* that X *believes that* $c \subseteq C$. If either of these beliefs is false – i.e. if Z does not have the authority to make claim R about Y, or if $c \not\subseteq C$ – X's trust will be misplaced as based on a misunderstanding. In any case, Y's interpretation of R in c may well be very different from X's.

In short, in definition (2) above, Z has to be such that X *believes that* he can authoritatively make claim R about Y, and I(R,c) is X's subjective interpretation of R within a set of contexts c, such that X *believes that* $c \subseteq C$.

---

[3] I shall use 'belief' as a basic noun covering the propositional attitude in question, but this is a linguistic shorthand. I do not wish to judge the philosophical issues of whether only humans can trust, or whether language is required for belief. The definition of trust given below is neutral between humans and non-humans. Can animals trust? Can organisations trust? Can artificial agents trust? Can babies trust? These questions boil down to the issue of whether they can hold an attitude toward another – *not* to the question of whether they are able to have beliefs.

Note that X has strategies available to reduce the uncertainties introduced by subjectivity. Most obviously, X can negotiate with Y to determine more precisely the content of Y's claim R, and how it will affect X in the contexts c with which he is concerned.

This leaves two more parameters. Deg is a measure of X's confidence in his attitude toward Y's trustworthiness. The metric for Deg depends on the system under discussion. For psychological realism, it may be that Deg would be a fairly coarse-grained Likert-type psychometric scale of five or seven points. But it would be legitimate to produce more complex models that modelled Deg on, say, the real line between 0 and 1.

Whatever metric chosen must facilitate two judgments that X will need to make. First of all, X may have to choose whether he trusts Y_1 more than Y_2 when he decides where to place his trust. Secondly, the level of risk that X takes on with respect to an interaction with Y will depend on his degree of trust; if he trusts her a lot, he will, all things being equal, be prepared to risk a lot, and if he trusts her only a little, his appetite for risk will be diminished.

Warr is the warrant for X's trust in Y. This could take any form – it doesn't have to be rational, and could even be that X has been dosed with oxytocin which increases the propensity to trust [17]. But usually there is a sensible rationale behind a trust judgment, which is important for assessing it, and also for assessing how robust it is likely to be. Typical relatively reliable trust warrants include the reputation of Y, the past history of X's encounters with Y, the availability of sanctions for X, the possibility of a binding reciprocal agreement between X and Y, the credible commitments made by Y and the credentials that Y brings to the transaction.

As Wierzbicki argues ([31], pp.26-27), trust that does not have a rational component will be hard to model. That does not mean that trust cannot be irrational, but it makes it harder to embed psychologically-realistic trusting mechanisms into software, or to design socio-technical systems (or social machines) which incorporate potentially irrational human trust judgments without restriction.

1.3. Failures of Trust and Trustworthiness

The problem of trust highlights the possibility of failure. The negations of (1) and (2) deserve some attention.

First of all, we should distinguish between a lack of trustworthiness (for some R and C) and untrustworthiness. There is an important gap between someone who makes no claims to trustworthiness and someone who makes a false claim. Barack Obama is not a trustworthy brain surgeon, but as we have already discussed this does not redound to his discredit; neither he nor any authoritative Z who speaks for him has made a claim to that effect. I am within my rights to point out that Obama is *not* a trustworthy brain surgeon, but I cannot use this as a criticism.

The term 'untrustworthiness' is reserved for those circumstances where Y *has* made a claim R which does not describe her intentions, capacities and motivations correctly. Even then, the level of discredit accruing to Y will vary. If she has deceived X deliberately, then she has behaved immorally and is deserving of censure if not proceedings for fraud. But her deception may be unintentional. For instance, she may have overestimated her capacity to perform R. She may have offered to help with her daughter's homework, but found herself incapable of remembering the mathematics of

her schooldays. In such cases, Y may be guilty of negligence, but not deliberate deception. However, in each case it is fair to describe Y as untrustworthy.

Failure to trust also comes in a variety of packages. What might make X fail to hold the appropriate attitude to Y as described by (2)? In the first place, he may never have heard of Y, or may be unaware of Y's intentions, capacities and motivations, and hence simply not have formed the appropriate attitude. Secondly, he may be aware that Y has made no relevant claim R (so he may not trust Obama to perform brain surgery, but that's OK as no-one has suggested he does).

The third option is active distrust, where X's attitude toward Y is fully informed (in X's opinion), and consists, in effect, in the belief that she is untrustworthy. In other words, X believes that Y's intentions, capacities and motivations are not as represented by R, which has been presented by an authoritative Z. This, of course, is a very strong claim that is worth more analysis than space here allows.

1.4. The Problem of Trust

The problem of trust, as argued above, is not to increase trust, but rather to ensure that X trusts Y when and only when Y is trustworthy. This is difficult as the incentives are not optimally aligned. If X risks assets in an interaction with Y, then he benefits from her *trustworthiness*, but unfortunately he only controls his *trust*. Similarly, Y benefits from X's trust, but only controls her trustworthiness. The result is a dilemma where the benefits of cooperation could be high, but losses to a trusting (trustworthy) party would accrue if their partner is untrustworthy (distrusting).

From this two things follow. First, trust cannot be an entirely rational attitude; it is not the sort of thing to survive rigorous game-theoretic analysis [14]. Second, X should use the analysis of (2) to determine where trust judgments can break down. Many failures of trust are down to differences in interpreting what Y is committed to.

A typical strategy for a trustworthy Y is to send *signals* of trustworthiness to X, which ideally will accurately represent her trustworthiness and which will be included in X's warrant to trust Y [25]. These signals can be conscious or unconscious, and more or less strongly connected with the task that Y is offering to carry out, preferably as an unavoidable by-product. The flip side of any such signalling system, however, is that if it is made explicit, then it can potentially be counterfeited by an untrustworthy person. Types of signal already mentioned include Y's reputation, history and credible commitments.

A second strategy involves structuring the encounter with some kind of *institution* which can reduce the likelihood of a deception being in Y's interest. Such an institution might supply credentials for Y, or might make plausible and effective sanctions available for X to apply if Y defects. Or X and Y might set up their own 'mini-institution' by entering into a reciprocal agreement.

If we generate some numbers, we can present the dilemmas for X and Y in a two-person one-shot game. Suppose there is a situation where, if X trusts Y with an investment of €3, each can walk away with €10 profit. X has the choice of trusting or distrusting Y; Y has the choice of being trustworthy or untrustworthy. We can assume that trustworthiness has costs of its own (say, €1), to cover the arrangements that Y would have to make to ensure that she can carry out her claims.

The situation then looks like this:

Table 1. Should X trust Y?

X\Y	Trustworthy	Untrustworthy
Trust	10,9	-3, 3
Distrust	0,-1	0,0

The upper left quadrant gives the maximum benefits of cooperation. If X trusts untrustworthy Y, then he is defrauded out of €3. If X fails to trust untrustworthy Y, nothing is gained or lost, while if X fails to trust trustworthy Y, we have opportunity costs for both parties, plus the expenses of trustworthiness for Y.

If X or Y is risk-averse, there is little hope of reaching the maximal payoff. X may reason that if he trusts Y, then, although he might gain €10, he might also lose €3. If he distrusts, then he can be no worse off than he is. Similarly, risk-averse Y might reason that if she is trustworthy, she might end up worse off than before, but if she is untrustworthy, she is no worse off and may even gain €3 from a credulous X.

2. Trust in the Digital World

The previous section was intended to be entirely general. As Bus has written, "the concepts of security, trust and as a consequence privacy as developed in democratic societies, are fundamental drivers for self-organisation in our society" [3]. Bus points out that the above theory can point X to particular strategies to investigate whether Y's untrustworthiness was due to a particular type of claim R, or a misleading agent Z. Such an investigation can lead to a reconfiguring of the relation between X, Y and Z (and alternatives for Z), and a renegotiation of the content of R and C.

The particular issue in the digital world is that the number of actors is larger, while the bandwidth along which they can send signals is lower, resulting in less information. The warrants for trust (particularly rational trust) are therefore being stretched, and a situation has been created where recognised signals for trust can be subverted. Phishing attacks, for instance, are dangerous not only for ordinary Web users, but even for those with high levels of computer literacy [4]. Complexity and uncertainty introduce greater vulnerabilities.

Bus adapts (1) to define the trustworthiness of technology as "the behaviour of the technology within a given context in conformity with its representation as published by the accredited agent" [3]. Y, on this reading, is the trustworthy system or technology. The accredited agent (Z, in our terminology of variables) consists of at least the software producer, dealer or marketing team, while R is a specification, advert or other representation of the technology's behaviour which is liable to influence X's decision to buy it or to rely upon it. These definitions will be expanded below.

First note that R cannot be restricted simply to the technical specification but must draw upon implicit factors which affect trustworthiness. It includes expectations of functionality of any software or technology, and also legal agreements and waivers, particularly privacy policies and anything else the user has consented to either implicitly or explicitly. Other factors include: robustness across environments; robustness against hostile attack; usability and the ability to cope with users' errors, compliance with law and regulation, such as data protection; and compliance with contractual agreements with users.

Note also an important shift between (1) and (2). The content of (1) is determined in part by Z's specification R of Y's behaviour and functionality, and the restriction of the scope of R to a set of contexts C (which could be done via the terms and conditions

to which the user consents by clicking a box). However, (2) depends on I(R,c), which is X's *interpretation* of R and its effects in a subset c of C (the subset of contexts which interest X).

It is clear that there is suddenly great scope for miscommunication and misunderstanding, particularly in the digital world where many claims are shrouded in technical obscurity, where there is great ignorance about the implications of the use of information, where there may be several legal jurisdictions covering a single transaction, and where there are relatively few widely-accepted norms of behaviour. In particular, it may well be that X's interpretation of R does not coincide with Y's, and that X's area of interest c is not a proper subset of C.

These issues can arise in particular when R is a very technical specification, but X's interpretation I is couched in the social terms which X understands. Bus's example of an ID card is a good one: the card may meet a complex technical specification R developed by a government Z, but X prefers to interpret the card in terms of his social and political interpretation of such issues as privacy, crime and security, and the disagreeable extension of powers of government that he has observed over the years. Z may be concerned with minimising legal liabilities, while X may be interested more in how he will be empowered and constrained by the new technology.

In other words, X and Z may simply be speaking different languages. A piece of technology is not seen as a secure piece of kit, but rather as an instrument of empowerment and/or security and/or repression, depending on the point of view of the trustor. Technical capabilities are not judged in the abstract. As Charles Raab puts it, "it is no comfort to a privacy-aware individual to be told that inaccurate, outdated, excessive and irrelevant data about her are encrypted and stored behind hacker-proof firewalls until put to use by (say) a credit-granting organization in making decisions about her" ([27], p.285).

The particular danger is that Z will focus, in the construction of R, on the technical specification in isolation. This, in the digital world, is an enormous mistake. X's interpretation of R will go way beyond the technical specification (and indeed from X's point of view the specification may be of very minor importance). X will focus on the policies to which he has consented (and, as we know that very few people scrutinise terms and conditions and privacy policies in detail, X's beliefs about those will be general, abstract and impressionistic), the marketing which has attracted him, and the affordances of the technology that he has observed in the world around him.

3. Enlightenment Thoughts on Trustworthiness and Trust

The complexity and uncertainty described by Bus and others exacerbate the difficulties caused by the subjective shift between (1) and (2), from Z's understanding of Y's intentions, capacities and motivations to X's. Yet as we work toward bringing about a Digital Enlightenment, it is worth noting that trust was a major concern of the original Enlightenment. The *philosophes* of that era wrestled with the question of how we ensure, or promote, trustworthiness in a world where untrustworthiness often pays. As David Hume put it, can we find "a remedy, in the judgement and understanding, for what is irregular and incommodious in the affections" ([15], p.489)? In particular, some of the great political thinkers of the 17th and 18th centuries can still give us important pointers as to how to deal with the conundrums discussed in section 2. Their work can also alert us to some of the pitfalls.

3.1. Hobbes and Leviathan

The most obvious point about promoting trust was made by Thomas Hobbes, who argued in *Leviathan* (1651) that an unregulated free-for-all would be undermined by the tendency of people driven by glory (or kudos), competition and the needs of the self to act against the common good. In conditions of resource-scarcity, the sum of our actions will defeat our individual ends. There is therefore a need for a sovereign, Leviathan (a state or government), to regulate and keep the peace [13].

Regulation will certainly be an important part of the story of how we link trust and trustworthiness. By providing constraints and sanctions on Y, Y's intentions, capacities and motivations are altered to make divergence from R less likely, and therefore X's degree of confidence in her trustworthiness should increase. Yet Hobbes' Leviathan has duties of its own, especially the maintenance of peace and security and the prevention of discord. Hobbes pioneered the idea of contractual bases for political settlements.

Hence Hobbesian regulation must be relatively lightweight, to allow citizens maximum liberty consistent with the prevention of discord, because that is the contractual basis of the relationship between citizens and Leviathan. This is especially important in the online world, given the extraordinary innovation that the Web has fostered in the last couple of decades.

Much commentary on Hobbes has focused on Leviathan's monopoly of power and legitimate violence; this monopoly is less relevant online, where the W3C's voluntary standards compete with various national and supranational entities asserting their power, and where code can regulate action just as easily as governments and standards setters [19]. But the contract between citizen and Leviathan does remind us of the many contractual relationships between Web service providers and users. At the moment, providers make the pace in such contracts. Many commentators have pointed out the unsatisfactory nature of our clicking consent boxes, and 'accepting' privacy policies by entering a website. This is absurd, and does not help establish warranted relations of trust between users and providers.

Regulation needs to be tailored more closely to the needs of Web users, in order to allow users to make informed decisions about consent etc. It is a common assumption that users are ignorant of technical matters (which may be true), and that therefore their decisions are uninformed. Nevertheless it is also true that decisions are framed in terms which make no sense to users.

For example, consider a privacy policy on a social networking site. This is an extremely important document for a user, because it may well affect his or her self-determination, if information or images from such a site become available to potential employers, family members or intimate friends indefinitely into the future. This is a complex issue, and the document itself will also be complex. Yet the act of acceptance is a crude binary decision. The decision is binding even if the policy changes in the future (which is doubly absurd). Furthermore, in the US at least, a company's privacy policy is regulated by the Federal Trade Commission which looks for deceptive or unfair practices – in other words, it asks whether the company did what it said it would do (which is an important part of the definition of trustworthiness). However, the FTC does *not* ask whether this specification accords to some independent definition of privacy. In other words, the FTC does not care whether a privacy policy actually protects users' privacy, as long as the company adheres to it.

In such cases, the company may be trustworthy by its own definitions, but may fail to instil trust in its users (i.e. Y conforms to R, but fails to conform to I(R,c)). Hobbes

reminds us of the contractual nature of a regulatory structure, which must therefore be sensitive to the needs of users as well as service providers. Consent, in particular, can and should be handled more sensitively.[4]

3.2. Burke and the Wisdom of Crowds

Regulation can never be the whole story with respect to trust. As the Enlightenment historian Edward Gibbon wrote, "a thousand quarrels must arise under a law, and among men, whose sole umpire was the sword." If regulation was the only constraint then innovative firms would circumvent it, or be untrustworthy when they were sure they would not be found out. Fortunately, however, the Enlightenment bequeathed more social tools and relationships which we can adapt as we build the Digital Enlightenment.

A number of Enlightenment thinkers investigated ways in which rational individuals sustain strong connections with their embedding societies, and these political discussions remain the basis of political philosophy today ([22], pp.66-114). We can make an immediate reference to the great conservative philosopher Edmund Burke, who pitched himself as an opponent of Enlightenment, but, I as have argued elsewhere ([22], pp.105-107), was a definite product of Enlightenment thought. Like Rousseau, he reacted against many of his contemporaries, yet produced a defiantly and clearly modern philosophy.

Burke's insight was that reason and design are not enough. Central committees and planners cannot create usable institutions. The most powerful institutions emerge from behaviour, and from practice that is meaningful to people. No small number of thinkers however well-intentioned and incorruptible can create a public space that has the same connections with citizens. Yet "Your literary men, and your politicians, and … the whole clan of the enlightened among us, … have no respect for the wisdom of others; but they pay it off by a very full measure of confidence in their own" ([2], p.184). Burke instead approved of what is now called the wisdom of crowds [30], and looked to tradition and even prejudice as ways of ensuring that society cohered – the effective aggregation of very many individual subjective points of view. The individual is in no position to judge wider social requirements. "We are afraid to put men to live and trade each on his own private stock of reason; because we suspect that this stock in each man is small, and that the individuals would be better to avail themselves of the general bank and capital of nations, and of ages" ([2] p.183).

It so happens, of course, that the Web has always included impressive methods to aggregate opinion, from PageRank to Wikipedia to recommender systems. Social networking has been an entirely unpredicted success, which has brought the Web in a meaningful way to very many more people (it has relevance for their leisure and private lives, as well as being a valuable work tool). The lesson from Burke is the power of the Web 2.0 paradigm.

He also pointed out the need for measures to be taken to preserve important public goods (for example, ensuring data protection), and to prevent what has been called the noosphere undermining people's individuality. It is essential to realise that system designers cannot decide for themselves how a system will be used; if it is used widely

[4] See the projects EnCoRe (http://www.encore-project.info/) and VOME (http://www.vome.org.uk/) which are beginning the task of providing more sensitive mechanisms for managing consent.

enough and is meaningful to enough people, its functionality will be a negotiation between users, administrators and designers.

3.3. Rousseau and the General Will

Such ideas get us closer to trust and trustworthiness, because the interactions that Web 2.0 supports are voluntary, and basically focused around common pursuits that help bind communities. Of course these can be undermined by individual impersonators who introduce bad faith, such as the American PhD student at a Scottish university (ironically, Adam Smith's and David Hume's old university) who posed as a lesbian Syrian blogger. They can also be undermined by more structural intercessions; for example, a social network might exploit the data it holds about its users to support targeted and intrusive marketing. This could undermine trust in the platform, which in turn might lead to a decline in the number of interactions occurring on it, although it need not necessarily reduce offline trust within communities of users.

In this context, Burke's ideas can be supplemented with additional thoughts from his philosophical opponent Rousseau (Burke opposed the French revolution, while Rousseau was Robespierre's chief source of inspiration). But they agreed on the importance of community. Rousseau theorised the *general will*, a democratic construct that viewed the interests of the people as a whole as something that transcended the interests of individuals or of particular factions. Rousseau makes it clear that "there is often a great deal of difference between the will of all and the general will; the latter considers only the common interest, the former considers private interest, and is no more than a sum of particular wills" ([28], p.203). Genuine aggregation of opinion is important. A legitimate consensus is formed when all individuals are treated equally, in an absence of factions ([28], p.225).

However, it is equally important that individuality is retained within, and against, consensus where appropriate [18]. Rousseau (unlike Burke) was not very sensitive to the risks of totalitarianism ([21], pp.109-125). Modern commentators such as Jaron Lanier emphasise the importance of retaining individuality within frameworks that allow liberty of expression. Rousseau's ambivalence about oppression should not, of course, be replicated today.

What does this mean for trustworthiness in the online world? In particular, it suggests that we should conceive of agency holistically. Individuals' preferences are of course important, but the public space of the Web is an important good that needs to be protected [7], [8], an argument that is already familiar in the context of the Web [32]. But whose agency matters?

Trust is an individual's judgment. Someone creating a trustworthy system certainly needs to think in abstract terms about its users, but from our 21st century viewpoint, it is to be hoped that individuals can be disaggregated to express themselves as individuals [18]. However, there is a point to aggregation. When we look at (1) above and consider the arguments at the end of section 3.2, it is clear that Y and Z cannot be treated as simple individuals to be judged within a framework that they themselves set; online trust, as noted, goes well beyond checking that a system meets a target specification.

Indeed, the system itself – represented by the variable Y in (1) – can be seen at several levels. The technology does not come unaccompanied, but is integrated into a social system with which the trustor is confronted. Particular individuals in that system, including the technological agent Y, are not the focus of judgments by trustors. The *system itself* is the important agent here, not the smaller technological component. We

should ensure, when considering online trust, that the aggregated system is seen as the important agent when trustworthiness is being designed and planned for, because that is the agent that X will be making his trust judgments about.

Similar thoughts apply to Z. This is not the system designer or producer in isolation; several individuals will have been involved, in any major online development, in creating a system using regulations and financial incentives as well as the actual code. Claims of trustworthiness of the form (1) should ensure that Y and Z are conceived as holistically as possible, because that is how X will be making his trust judgments. If the interests of Z and X are to be aligned (so that trustworthiness and trust can flourish simultaneously), they have to agree on the object of X's trust.

Furthermore, we must ensure that any system design does not privilege anyone's interests, and does not assume that the system is the only thing at issue. Annette Baier complains that "the more we ignore dependency relations between those grossly unequal in power and ignore what cannot be spelled out in explicit acknowledgement, the more readily will we assume that everything that needs to be understood about trust and trustworthiness can be grasped by looking at the morality of contract" ([1], p.106). Baier's point is highlighted when we compare Hardin's definition of encapsulated trust ([10], pp.16-20) with Hayek's definition of coercion ([11], p.133), and discover that they are extremely similar. Trust can become coercive if we are not careful (for example, a social network might allow users to express themselves, while insisting they do so in circumscribed ways). Burke's and Rousseau's work reminds us of the importance of social units over and above their component individuals, and of the importance of individual liberty in the context of environments (such as the Web) which give meaning to liberties while simultaneously limiting behaviour. That is not to say that either man resolved the many tensions such issues raise.

3.4. Hume and Sympathy

David Hume also extended the Enlightenment toolkit in this area by developing the idea of *sympathy*. We are driven not simply by our personal desires, or by the needs of our societies and communities; we have a hard-wired interest – indeed pleasure – in the well-being of our fellow-people [15]. Because of this, their narrow interests become our wider interests.

This feeds into the development of trustworthy online systems via a sympathetic understanding of C, the set of contexts in which Y is claimed to be trustworthy. This set hopefully is a superset of c, the set of contexts in which X has an interest. Often there is a mismatch here: the terms and conditions to which X has agreed may rule out Y's function in certain important contexts in c. However, terms and conditions are often very long, in legal language, and intended to reduce Z's liabilities; they are very rarely in place to help X. The result is that X does not internalise them, and may assume that Y will function to his benefit in contexts where it is not intended to. The result is a loss of X's trust. If Z had a sympathetic understanding of X's requirements and preferences, then Z would try to ensure that Y functioned when X wanted it to – not when it was convenient for Z.

Hume tells us, when we design systems, to establish a set of contexts C which cover the multitude of contexts in which a user might be interested. This means that C will be more heterogeneous and complex, and less easy to control. However, by ensuring that trustworthiness is not narrowly confined to well-understood contexts, X's trust is more likely to be earned and warranted.

3.5. Smith and the Impartial Observer

Adam Smith's idea of the impartial observer has a similar effect to Humean sympathy. Smith extends Hume's account. Sympathy restrains our selfishness; however, Smith also contended that we are able to use the sympathy we have for others to develop standards of behaviour to enable society to function more smoothly and in the interests of all. These standards are internalised as an "impartial spectator" [29], an idealised person who acts as a model. People act as if their actions were being observed and judged by the impartial spectator, and the extent of their wisdom and virtue depend on how often they follow the imaginary spectator's 'suggestions' in opposition to whatever is dictated by their own desires. The man of wisdom "almost identifies himself with, he almost becomes himself that impartial spectator, and scarce even feels but as that great arbiter of his conduct directs him to feel" ([29], p.147).

Hence someone hoping to design trustworthy systems needs not only to understand the contexts that users are interested in as we noted above, but also needs to create holistic system behaviour that will meet the needs of users. In the terms of (1), R needs to be tailored to the complexity reflected in a holistic interpretation of C, but also needs to specify behaviour that enables the system to address the needs of users holistically conceived.

Hume's idea of sympathy enabled Smith to derive the closely related but more objective notion of the impartial observer; understanding what matters to individuals can lead us to conceptualise the behaviours that they will find helpful and useful. Similarly, rethinking the contexts in which a trustworthy system operates should enable system designers to create systems that function in such contexts.

3.6. Kant and the Categorical Imperative

The final Enlightenment thinker that we can learn from about issues of trust and trustworthiness is Kant, and in particular his conception of the *categorical imperative*, which states that moral laws must be general ones [16]. For our purposes, the important corollary is that people should be treated as ends, not means. Trustworthy systems should serve their users, rather than treat them as resources to be exploited.

So, to take an example, a social networking platform which encourages people to place data upon it primarily in order to harvest that data and use it for marketing purposes will inevitably lose trust. That is not to say that the data cannot be harvested, only that the primary purpose of the platform ought to be the facilitation of interaction, and a good user experience. Similarly, a search engine whose primary purpose was targeted advertising will lose trust, especially if users begin to suspect that the search results were being skewed as well; the primary purpose of the engine should be to present the most useful web pages in the most useful order given the query. If, once that purpose is accomplished, it proves possible to generate valuable services on the back of it, trust is more likely to be warranted and preserved.

Kant's discussion is in terms of the duties of a moral actor. However, the basic principle of not being self-centred, but being *other-directed* is the important part. It also connects with Smith's notion of the market and its role in promoting interaction and trust ([23], pp.221-224). Regulation treats the individual as a self-centred actor, while markets ensure that a more sociable view prevails; if I wish to flourish in a market, I must provide products and services that users want. I need to take those users' preferences seriously, and that includes developing trustworthy products along the lines

that I have discussed in this paper, holistically conceived and worthy of warranted trust from users.

And so we move full circle; regulation has an important place in ensuring that the Digital Enlightenment provides an important public space for interaction between citizens, but the private sector will always have a vital role, both as a force for innovation, and a guarantor of trust.

4. Conclusion: Hints From the Enlightenment

Of course, it is obvious that this paper is not intended as an in-depth discussion of the Enlightenment thinkers, or as an original contribution to historical scholarship. The arguments are analogies, helping us develop a roadmap, rather than a fixed plan of action. However, I hope it shows how the thinkers of a previous time can still teach us lessons about social arrangements in the digital world – in broad terms at least. We should not be surprised at this – as many have argued, the political philosophy of the Enlightenment sets out the basic lines of political discussion that we still recognise today [14], [22].

In this paper, I have set out a theory of trust, and shown how we can interpret a number of issues surrounding trustworthy systems on line using that theory as reference point. The Enlightenment thinkers were enlisted to show how issues of trustworthiness were dealt with in earlier times – the connections *between* people, in terms both of basic hard-wired attitudes and of institutional connections, which enable people to cooperate and interact without exposing themselves to too much risk. This was the key property of the public space that emerged in the 18th century [7], and is also a key desideratum of the World Wide Web [32].

If we think about the ground we covered between Hobbes and Kant, two key factors emerge. First, we need to ensure that systems designers take a holistic view of the relevant agents, contexts and behavioural claims made, as Bus has argued [3]. Trust can only connect with trustworthiness if the trustor and trustee speak to each other in the same language. X will naturally see a system in a wider context, while Z may be tempted to retreat to a narrow technical view. There are severe opportunity costs for Z if trustworthy systems are not trusted.

Second, designers need to be other-directed. Exploitation of trusting users as a primary function of a system will eventually undermine trust. This will be an important point particularly for a new breed of 'social machines' integrating human and machine problem-solving [12]. Any such integration must serve the purposes of the humans in the loop. As Jaron Lanier has recently written in a powerful polemic, "if you get in deep enough, you get trapped. Stop calling yourself a user. You are being used" ([18], p.200).

Acknowledgements

The work reported in this paper was funded by the EnAKTing project, EPSRC Grant EP/G008493/1.

References

[1] A.C. Baier, *Trust and antitrust*, in *Moral Prejudices: Essays on Ethics*, Harvard University Press, Cambridge MA, 1994 (1986), 95-129.
[2] E. Burke, *Reflections on the Revolution in France*, Penguin, Harmondsworth, 1968 (1790).
[3] J. Bus, *Trust, self-organisation and complexity in digital space*, This volume, IOS Press (2012).
[4] R. Dhamija, J.D. Tygar & M. Hearst, *Why phishing works*, Conference on Human Factors in Computing Systems (CHI 2006): http://www.cs.berkeley.edu/~tygar/papers/Phishing/why_phishing_works.pdf.
[5] D. Diderot, Encyclopédie, in I. Kramnick (ed.), *The Portable Enlightenment Reader*, Penguin, New York, 1995 (1755), 17-21.
[6] F. Fukuyama, *Trust: The Social Virtues and the Creation of Prosperity*, Basic Books, New York, 1995.
[7] J. Habermas, *The Structural Transformation of the Public Sphere*, Polity Press, Cambridge, 1989 (1962).
[8] J. Habermas, *The Theory of Communicative Action Vol.1: Reason and the Rationalization of Society*, Polity Press, Cambridge, 1984 (1981).
[9] R. Hardin, *Trustworthiness*, Ethics **107** (1996), 26-42.
[10] R. Hardin, *Trust*, Polity Press, Cambridge, 2006.
[11] F.A. Hayek, *The Constitution of Liberty*, Routledge, London, 1960.
[12] J. Hendler & T. Berners-Lee, From the Semantic Web to social machines: a research challenge for AI on the World Wide Web, *Artificial Intelligence* **174** (2010), 156-161.
[13] T. Hobbes, *Leviathan*, Penguin, Harmondsworth, 1968 (1651).
[14] M. Hollis, *Trust Within Reason*, Cambridge University Press, Cambridge, 1998.
[15] D. Hume, *A Treatise on Human Nature*, Oxford University Press, Oxford, 1978 (1740).
[16] I. Kant, *Groundwork of the Metaphysics of Morals*, Cambridge University Press, Cambridge, 1997 (1785).
[17] M. Kosfeld, M. Heinrichs, P.J. Zak, U. Fischbacher & E. Fehr, Oxytocin increases trust in humans, *Nature* **435** (2005), 673-676.
[18] J. Lanier, *You Are Not a Gadget*, Penguin, London, 2011.
[19] L. Lessig, *Code: And Other Laws of Cyberspace*, Basic Books, New York, 1999.
[20] N. Luhmann, *Trust and Power*, Wiley, New York, 1979.
[21] T. O'Hagen, *Rousseau*, Routledge, London, 1999.
[22] K. O'Hara, *The Enlightenment: A Beginner's Guide*, Oneworld, Oxford, 2010.
[23] K. O'Hara, *Conservatism*, Reaktion, London, 2011.
[24] K. O'Hara, *A general definition of trust*, working paper (2012), http://eprints.ecs.soton.ac.uk/23193/.
[25] A. Pentland, *Honest Signals: How They Shape Our World*, MIT Press, Cambridge MA, 2008.
[26] P. Pettit, Trust, reliance and the Internet, *Analyse & Kritik* **26** (2003), 108-121.
[27] C.D. Raab, The future of privacy protection, in R. Mansell & B.S. Collins (eds.), *Trust and Crime in Information Societies*, Edward Elgar Publishing, Cheltenham:, 2005, 282-318.
[28] J.-J. Rousseau, *The social contract*, in G.D.H. Cole (ed.), *The Social Contract and Discourses*, J.M. Dent, London, 1993 (1762), 179-309.
[29] A. Smith, *The Theory of Moral Sentiments*, Liberty Fund, Indianapolis, 1976 (1759).
[30] J. Surowieki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*, Little, Brown, London, 2004.
[31] A. Wierzbicki, *Trust and Fairness in Open, Distributed Systems*, Springer, Berlin, 2010.
[32] J. Zittrain, *The Future of the Internet: And How to Stop It*, Penguin, London, 2008.