Ignore These At Your Peril: Ten principles for trust design

Jens Riegelsberger, Google
M. Angela Sasse, UCL

ABSTRACT

Online trust has been discussed for more than 10 years, yet little practical guidance has emerged that has proven to be applicable across contexts or useful in the long run. 'Trustworthy UI design guidelines' created in the late 90ies to address the then big question of online trust: how to get shoppers online, are now happily employed by people preparing phishing scams. In this paper we summarize, in practical terms, a conceptual framework for online trust we've established in 2005. Because of its abstract nature it is still useful as a lens through which to view the current big questions of the online trust debate - large focused on usable security and phishing attacks. We then deduct practical 10 rules for providing effective trust support to help practitioners and researchers of usable security.

1. Introduction

Over the past ten years, 'trust online' has been the topic of research and debate in academia and industry, and even in the popular press. It is interesting to note that - while the same terms are still being used - the issues and underlying risks being discussed have shifted. In the late nineties, the key concern was a 'lack of trust in online interactions', with prospective e-commerce vendors worrying whether would-be customers would be willing to enter credit card details and other personal information on the Internet. Most of the academic research and commercial advice published then focussed on 'how to increase user trust online' by making websites 'user-friendly' and having a 'personal touch' e.g. in the form of photos of company staff. Unfortunately, much this advice on how to make your Internet presence trustworthy is now being used by perpetrators of phishing scams, who are using the latest 'trustworthy UI design techniques' to trick users into revealing authentication credentials and other personal data. A key trust issue that has emerged with the huge popularity of social networking is users' voluntary (and sometimes ill-judged) disclosure of personal information, and accidental sharing of that data by applications and other users. Other key issues today are trust in co-workers in remote collaboration settings, in fellow players in online games, in e-government and electronic voting systems. In each of these areas, the actors and areas of risk involved vary considerably.

Given the variety and speed with which these different application areas have emerged, it is perhaps not surprising that that not much progress has been made in establishing unifying model that explains how trust in

mediated interactions is established, and how it can be fostered through design of systems and the ecologies that surround them. As empirical research on online trust tends to explore the issues in one domain, the results often cannot be generalised to other systems and application areas. In our view, little substantive progress has been made in our understanding of trust issues, due to the absence a coherent framework of what trust is and how it works, and the diverse terminology used by different researchers.

The aim of this position paper is to outline such a unifying framework to pull together research and discussion around online trust in different disciplines. We have noticed some recent systems and research efforts where we feel that use of the framework would provided a better understanding of the issues, and helped to avoid mistakes. This position paper is an attempt to present the framework in a more accessible format, and demonstrate how system designers and researchers who are active in this field today can use the framework through 10 practical rules. We apply these rules to include practical challenges drawn from the current debate around trust online: phishing, social networking, reputation ratings.

2. Trust

The term *trust* is used widely in everyday language, and its meaning varies considerably across different contexts. Its use in the scientific community, is unfortunately, not very different. Trust has been studied for many years in many disciplines, and there is a plethora of trust definitions researchers can choose from (Corritore, Kracher, and Wiedenbeck, 2003b). The definitions contrast on various dimensions and consider trust in different situational contexts. While there is no agreement on the definition of trust, and how to measure it, the importance of trust is rarely disputed.

The key function of trust is that it enables transactions that could otherwise happen, because the risks associated with the transaction would be perceived as too high by one or both parties. Since modern life is defined by a high dependency on others' actions, trust underlies most of our activities (Giddens, 1990). In a functioning high-trust enviroment, all participants reap economic benefits, because giving up direct control of assurance processes (such as vetting every transaction partner, and his/her history) frees resources for activities which we are more productive. To put is simply: assurance activities (creating rules, monitoring behaviour, enforcing sanctions) cost time and effort, which is taken away from primary production tasks.

Most researchers agree that trust is required in situations where an individual takes a risk, and there is uncertainty about the outcome (Mayer, Davis, and Schoorman, 1995; Luhmann, 1979).Uncertainty arises from the fact that future actions of others cannot predicted with absolute certainty.

3. Framework for trust in mediated interactons

3.1 The Basic Model

We develop the framework from the sequential interaction between two actors, the *trustor* (trusting actor) and the *trustee* (trusted actor) – e.g. a human or an e-commerce vendor and its technology.

Figure 1 shows a model of such a situation (Berg et al, 2003, Bacharach & Gambetta, 2003). Both actors can makesome gains by making a transaction. Before the transaction, trustor and trustee perceive signals (1) from each other, and the context of the transaction. The trustor's level of trust will be influenced by the signals perceived from the trustee and context. Depending on her level of trust and other factors (e.g. the availability of outside options), the trustor will either engage in trusting action (2a), or withdraw from the situation (2b) . *Trusting action* is defined as a *behaviour that increases the vulnerability of the trustor* (Corritore et al., 2004).

According to this model, a trustor will engage in trusting action if she perceives her potential gain - that will be realised if the trustee fulfills his part of the exchange (3a) - to be worth taking the risk. The risk is that trustee may lack the *motivation* to fulfill as promised, and decide to exploit the trustor's vulnerability, or he might simply not have the *ability* to deliver what the trustee expects (Deutsch, 1958). The trustee may not realise that what he intends to deliver is not what the trustor expects, or intentionally misrepresent his side of the bargin. Both possibilities result in non-fulfillment (3b).

In the absence of any other motivating factors, being trusted and then refusing to fulfill (3b) would be the logical outcome. However, in most real-world situations, we observe trusting actions and fulfillment despite incentives to the contrary: vendors deliver goods after having received payment, banks pay out savings, individuals do not sell their friends' phone numbers to direct marketers. In fact in many inherently risky situations we see trust being given, often habitually. Why is this? In many everyday interactions, trustee and trustor act under the influence of *trust-warranting properties* (Bacharach and Gambetta, 2003), i.e. intrinsic or contextual factors that provide incentives for fulfilment. Identifying and reliably signalling trust-warranting properties is the key concern for the emergence of trust and trustworthy behaviour.

If a trustee had accurate insight into the trustee's reasoning or functioning, there would be no need for trust (Giddens, 1990). Uncertainty - and thus the need for trust - stems from the lack of detailed knowledge about the trustee's trust-warranting properties. Information about these is only available in the form of signals (1). If trustor and trustee are separated in space, their interactions are mediated (e.g. by mail, email, telephone), and some of the signals that are present in face-to-face encounters may not be available or become distorted.

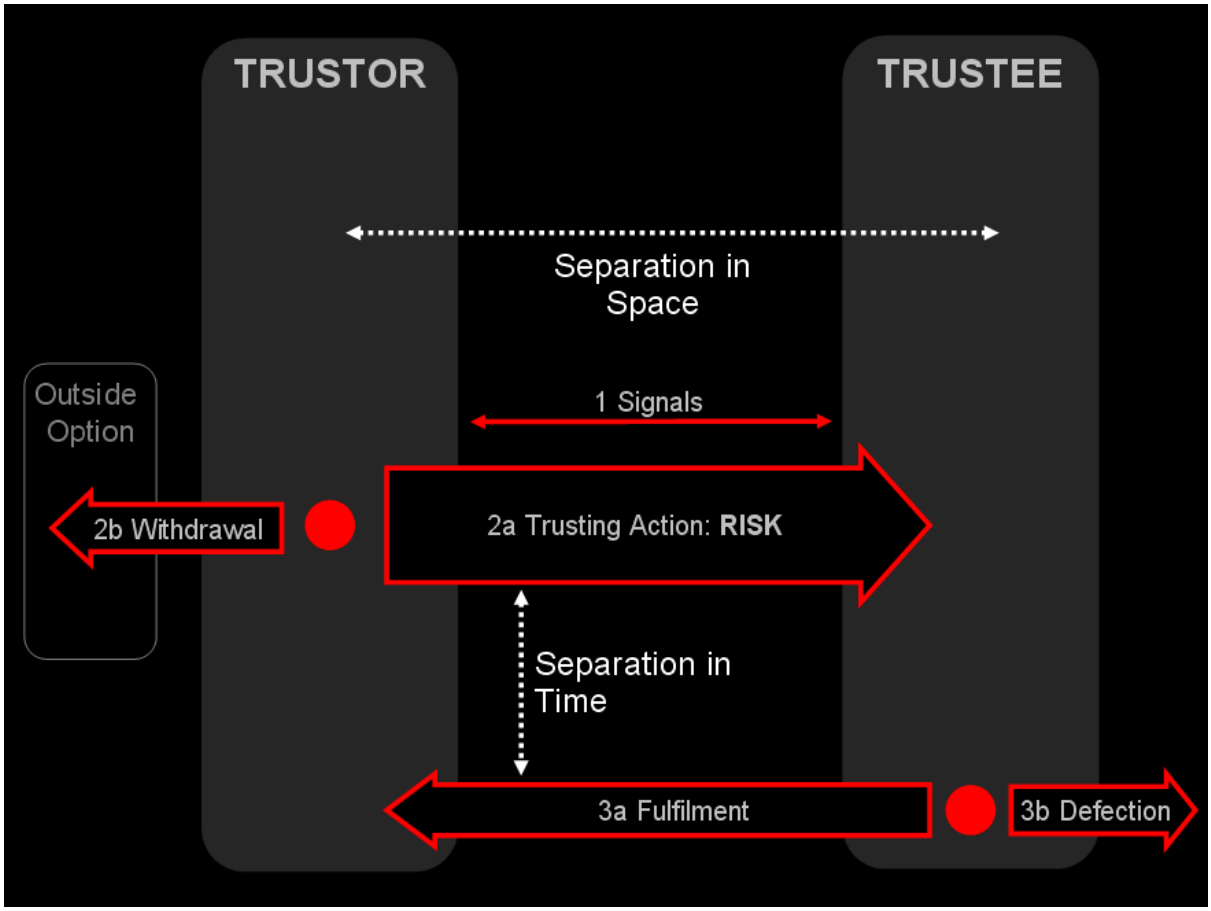

Figure 1. The trust-requiring situation.

3.2 Signalling Trustworthiness: Symbols and Symptoms

Failure of trust can also be seen as result of *mimicry* (Bacharach and Gambetta, 2003): non-trustworthy actors try to *appear trustworthy,* to induce the trustee to trust, even though they have no intention of fulfilling. To understand how mimicry can operate we draw on semiotics, and distinguish between two types of signals: symbols and symptoms (Riegelsberger, Sasse, and McCarthy, 2003b; Bacharach and Gambetta, 2003).

Symbols of trustworthiness. Symbols have an arbitrarily assigned meaning - they are specifically created to signify the presence of trust-warranting properties. Examples of symbols for such properties are e-commerce trust seals. Symbols can be protected by making them very difficult to forge, or by threatening sanctions in the case of misuse. They are a common way of signalling trustworthiness, but their usability is often limited. Because they are created for specific settings, the trustor has to know about their existence and how to decode them. At the same time, trustees need to invest in emitting them and in getting them known (Bacharach and Gambetta, 1997).

Symptoms of trustworthiness. Symptoms are not specifically created to signal trust-warranting properties; rather, they are by-products of the activities of trustworthy actors. As an example, a steady gaze and firm voice

may not require much effort when telling the truth, but may require some training to maintain while lying. Therefore, exhibiting symptoms of trust incurs no cost for trustworthy actors, whereas untrustworthy actors would have to invest some effort to engage in effective mimicry.

3.3 Trust-Warranting Properties

Having established the basic terminology, we introduce the factors that support trustworthy behaviour in transactions.

3.3.1 Contextual Properties

Raub and Weesie (2000b) identified three categories of factors that can lead trustees to fulfill. These are *temporal*, *social* and *institutional embeddedness* (see Figure 2).

Temporal embeddedness. If trustees have reason to believe that they will interact again with a given trustor (expectation of continuity) where they are recognizable (i.e. have stable identities), they have an incentive to fulfilll: while a trustee could gain through 'take the money and run', he also knows that the trustor would not place trust in future encounters. Non-fulfillment in the present encounter thus prevents gains that could be realised in future exchanges (Friedman, 1977; Axelrod, 1980).

Social embeddedness. Reputation is historic information about trustors' attributes, such as *honesty*, *reliability*, or *dependability* (McKnight and Chervany, 2000; Corritore et al., 2003; Friedman, Kahn, and Howe, 2000; see 2.3.2). Assuming these attributes are stable across time and context, they can form the basis of trust in present encounters. Reputation also has a second function, because of the ability of a socially well-embeded trustor to tarnish the trustee's reputation. It thus provides another incentive to fulfill.

Institutional Embeddedness. Institutions are organisations that influence the behaviour of individuals, or other organisations, towards fulfillment. Examples of institutions are law enforcement agencies, judicial systems, trade organisations, or companies. Institutions are often embedded in wider networks of trust where one institution acts as guardian of trust for another one (Shapiro, 1987). If fulfillment is assured only or mostly by institutional embedding, we typically speak of assurance, rather than trust. Institutional controls typically carry costs in terms of investigation and sanctioning processes.

In summary, contextual properties provide incentives for the trustee to behave in a trustworthy manner. Their presence allows trustors to engage in trusting action without detailed knowledge of the trustee. However, when trust is solely based on these properties – as is the case in compliance-based on assurance mechanisms – it is likely to break down in their absence.

3.3.2 Intrinsic Properties

While contextual properties can motivate trustees to fulfill, they do not fully explain how actors behave in the real world (Riegelsberger et al., 2003b). Contextual properties are complemented by intrinsic properties, which we define as relatively stable attributes of a trustee (Deci, 1992).

Ability. This property is the counterpart to motivation in Deutsch's (1958) classic definition of trustworthiness. Mayer, Davis, and Schoorman (1995) define ability for human and institutional actors as a *"... group of skills, competencies, and characteristics that enable a party to have influence within some specific domain"* (p. 717). Ability also applies to technical systems in the form of *confidentiality*, *integrity (accuracy, reliability), authentication*, *non-repudiation*, *access-control*, and *availability* (Ratnasingam and Pavlou, 2004).

Internalized norms. Granovetter's (1985) classic example of the economist, who – against all economic rationality – leaves a tip in a roadside restaurant, even though he never expects to visit it again illustrates the effect of *internalized norms*. For most people, compliance with norms is be internalized to such an extent that it becomes habitual. However, social norms differ across groups and cultures - they have to evolve over time, and to trigger them, the trustor has to signal his membership of the relevant group clearly (Fukuyama, 1999). However, not all norms induce trustworthy behaviour. And when they do, they can be exploited - for instance, claims of authority or pleas for help are often used to manipulate victims of social engineering attacks and phishing scams (see below).

Benevolence. The intrinsic property *benevolence* captures the trustee's gratification from the trustor's well-being. Hence, it is different from the expectation of future returns. The capacity for *benevolence* is an attribute of the trustee, but the specific level of *benevolence* is an attribute of the relationship between trustor and trustee. A trustee may act benevolently towards one trustor, but not towards another one.

In summary, intrinsic properties provide motivation and the ability for trustworthy behaviour that is independent from contextual incentives. Figure 2 shows the framework, based on the abstract situation introduced in Figure 1, with contextual and intrinsic properties added.

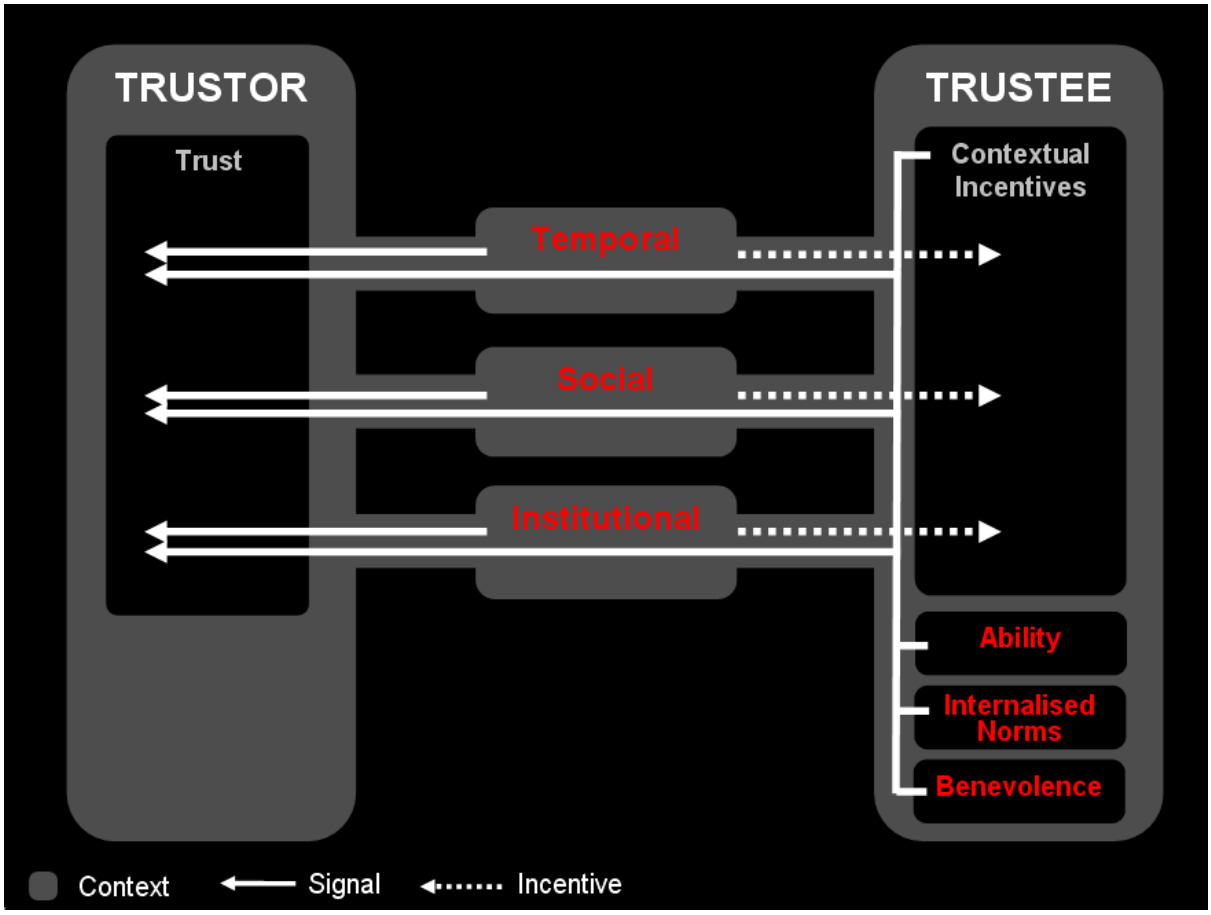

Figure 2. The complete framework.

4. Ten rules for providing effective trust support

1. Trust assessment is a secondary task

A trustor's assessment of the trustworthiness of a trustee does not take place in isolation - like all security tasks, this is a what in usability is called *secondary* or *enabling* task. Trustors initiate transactions with primary goal in mind - order the birthday gift that is delivered in time, to pay a credit card bill before incurring a penalty. Discharging the tasks required to reach these goals in a secure manner - i.e. a way that does not put your credentials at risk - is a secondary goal.

This focus on the primary goal drives user behaviour - trustors are focussed on the primary task that will allow them to achieve their goal. Security tasks are secondary tasks that interrupt users' main activity, and are therefore viewed as a necessary evil at best. Users want to spend as little time and attenion on as possible. Security mechanisms that demand users' attention and cogntive resources every time they interact with a system

are not only wasting their time and emotional energy, but will always be prone to attacks a) because users make errors in such situations, and b) because diverting users' attention is one of the most simple and effective attacks (Stajano & Wilson 2009).

There is another good reason why effective trustbuilding/security solutions must take users' primary goals into account - a security countermeasure that is effective in the some contexts will not work in others. In some contexts, trustors have to make very quick decisions on whether to engage in a transaction - for instance when buying tickets for events that sell out quickly. By the time you've checked up all possible security indicators (ULRs, padlocks, reputation scores, etc.) and established that the site is genuine, the tickets will be gone. Another issue that arises from goal-driven behaviour is that when trustors think they are presented with a something they really, really want, they want to believe the opportunity is a true one - even if it is to good to be thus. To be effective, security mechanisms need to effectively tackle the wishful belief. Doing so requires flagging explicitly that the trustor is not likely to get what they want. (See Principle7 for a more detailed discussion of the implications.)

The conclusion here is that, to be effective, the design of trust signals and security mechanisms need to support users' goals, and minimise interruptions and secondary workload.

2. Trust requires risk and uncertainty

As mentioned above, trust only happens in situations in which there is some level of *experienced uncertainty* regarding the outcome and where there is risk (i.e. the outcome has some value to the individual).

This means that researchers evaluating new systems in terms of their capacity to build trust, or signal trustworthiness appropriately, need to create risk and uncertainty for participants in their trials. For example, research evaluating the effectiveness of anti-phishing measures needs to have participants at least believe that they will lose money or their credentials if they place their trust in the wrong website. The effectiveness of an intervention in building trust or signaling trustworthiness cannot be assessed by having study participants just inspect a website, and ask them if they trust it. Conducting such evaluations without the presence of risk or uncertainty is akin to trying to evaluate the usability of a system by asking participants to look at it and say if they like it, without establishing that they are able to carry out essential tasks effectivly and efficiently.

An absence of risk, uncertainty, and also rewards (e.g. in the case of a phishing scam) will not create the incentive to in participants to engage in in-depth processing of cues they perceive on a website. This is well-known in empirical economics (e.g. Fehr & Fischbacher 2003) were studies often create considerable financial risk for participants - albeit without deception, and under the supervision of ethical review boards. However, it is often absent from research in usable security or security.

An illustration of the effect of the presence of risk and incentives on the evaluation of trust in experimental settings was given by Riegelsberger, Sasse and McCarthy (2005) in a study investigating participants' ability to detect cues of trustworthiness in an advisor presented in different media formats. The study was framed as a game, similar to 'Who wants to be a Millionaire', and the presence of even minimal risk led to an effect in users' behavior: they were more likely to rely on on advice by people represented through rich media representations (photos and videos).

Finally, researchers need to be aware that the emphasis has to be is on *experienced* uncertainty - not an actual one. If users engage in situations that carry risk, but are not aware of it, an evaluation of trustworthiness will not part of their cognitive processes. This is an aspect we discuss below.

3. Support *reliance*, as well as trust

After successfully interacting with a trustor several times - e.g buying goods from an online website - a trustor's view of her interactions change from the 'trust stance' described in 3.1, to one of *reliance*. The trustor now has an expectation that this trustee will deliver for this type of transaction. There is also a 'halo effect', meaning that they may regard the trustor generally as "trustworthy." This transition is key stage in the development of trust between two parties, and it is leveraged by many systems to provide trust signals for the whole community, for instance through reputation feedback and recommendations (see section 3.3.1).

However, the change from trust to reliance creates new vulnerabilities. In the inital stages of trust-building, the trustor is aware that she is vulnerable, and keeps a 'watching brief' to see if the trustor will exploit the vulnerability. If he does, she will not initiate another transaction. In e-commerce transactions or auction websites, most inexperienced users "dip their toe" into transacting online by making small purchases; once these have been successful, they will not expect problems with the same transactions in future. Many current attacks exploit this:

i. In phishing , most attacks impersonate websites the user is familiar with. Since security is a secondary task from the user's perspective (see principle1), and the whole point of building trust is to improve the efficiency of transactions, users will not check all details of the pages - such as security indicators. Instead, they respond to a small number of trust signals (see 3.2), such as the general appearance of the page, logos etc. This response is mostly automatic. Security experts tend to characterise this behaviour as 'careless', and advocate user education to check security indicators as a an effective countermeasure (e.g.Kumaraguru et al., 2009). Herley (2009), however, makes a convincing case that such countermeasures fail to understand basic economic principles that govern user behaviour - checking security indicators every time you go to a web page simply does not make for efficient interaction. To put it im terms of risk expert John Adams (Adams 1991), carrying out a risk assessment every time you cross the road would make for a low-risk, but rather unproductive life. These solutions also at odds with basic usability principles, since it diverts users' attention from their primary interaction goals (buy this birthday present before it is too late) and interrupts their primary tasks. An usable

and effective countermeasure should not interrupt users' task, and only draw their attention when a fraudulent/ suspect site is detected. The irony is that many current security solutions that aim to provide users with something 'they can trust' actually expect users to carry out laborious and time-consuming checks to make sure interaction is safe - basically, not to trust what they see without careful checking (meaning the efficiency gains of trust are not realised).

ii. After several successful transactions, the 'halo effect' kicks in, and the trustor starts to assume that the trustee is generally trustworthy. Trustees can exploit an existing trust relationship with an trustor by "cashing in" when trustors move from "dipping their toe" into small transactions to more substantial one. Security mechanims designed to leverage indiviudal trust relationships - such as reputation inidicators and recommendations - can also be exploited (see principle10).

In summary, it is important to distinguish between transactions governed by trust and reliance - something that is currently not done in research or commercial practice - move the focus from supporting users during the trustbuilding stages of interaction, to the more common reliance ones.

4. Identity *vs.* accountability

Identity is central to all trust-warranting properties discussed in section 3.3 , but in particular for the contextual ones ones in section 3.3.1 temporal, social, and institutional embedding. If actions cannot be reliably tied to trustors and trustees, the incentives from these contextual properties are not effective. If taking the money and running does not have consequences for the trustee's reputation, and thus reduce their prospects in future encounters, social embedding is no incentive for trustworthy behavior. Similarly, repeat transactions between the same trustor and trustee would not be recognized as such (there would be no effect of temporal embedding), and institutional enforcement is not possible if the malicious trustee cannot be identified.

Given that such much hinges on identity, it is not surprising that we have impressive identity checking capabilities: humans' ability to recognize and remember faces is astonishing, our brains are specifically attuned to the task - even to the extent that we can exploit the ability to recognize minor variations in face-shaped configurations to convey complex data in Chernoff Faces (1973). When humans started to live in bigger, more distributed communities, we were quick to devise systems to assure identity even among people who hadn't met before or were unlikely to meet again: seals, badges, uniforms, or - more recently - passports.

But relying on a single identities, and monolithic system to check them, creates a potential single point of failure and attack. It also creates privacy risks. If a sellers in an auction system, for instance, had to reveal their real-world names, anyone who knows them can see when they are putting the family silver up for sale. The main point of a stable identifier in trust-based transactions the ability to hold trustees accountable, and this does not require monolithic identity systems.

The conclusion here is that people manage their relationships through selective disclosure of information.

Designers must support them in their ability to do so, by providing multiple ways in which an identity can be expressed. Systems should provide users with a choice of how to when accountability is needed, and be able to control data associated with their identifiers.

5. Do you need trust - or assurance?

The economic benefits of successful high-trust environments have been demonstrated many times (see section 1). However, there are exchanges between trustors and trustee where the incentive structures are such that trust may not be the approriate framework for transactions. This is the case when the potential reward of defecting for the trustee is high - many trustees will be induced to 'take the money and run'. The old saying *'opportunity makes the thief'* has been confirmed by empirical research in criminology: most people will exploit opportunities that they come across if they perceive the risk of being caught as low. In online transactions, this would be the case if the trustee has little expectation of future transactions with a specific trustee, and the rewards from not fulfilling their side of the transaction are high. In such situations, designer should not aim to support trust, but systems that assure that trustees will follow a set of rules (Flechais, Riegelsberger, Sasse 2005). Institutional embeddeding (see 3.3.1) may provide some assurance with one country's legal frameworks, but will lack enforcement in trans-national exchanges (see Principle 9). More elaborate legal and technical solutions, such as trusted third parties and escrow services may be required.

One potential drawback of not using trust mechanisms in a system solution is that signalling to someone that they are not trusted may reduce their motivation to act in a trustworthy manner. If any previously unknown vulnerabilities emerge in a system, a trustee's motivation to exploit them is likely to be increased. To prevent this, designers can emphasise the institutional nature of assurance mechanism , communicating why it is necessary to follow certain rules, rather than leaving the trustee to feel that they are not trusted: *"It's business, not personal"*

The conclusion here is that designers need to understand the difference between transaction that can be supported by trust, and those that require assurance. The economics of transactions carried out under each scheme are different, and different design strategies are required.

6. Beware the power of social norms

As we established Section 3.3, there are broadly-speaking two types of incentives to act in a trustworthy fashion: contextual and intrinsic properties. Probably the most important intrinsic property are *internalised norms*: patterns of behaviour that we almost automatically engage in, as they have been instilled in us from an early age, through a process known as socialisation. Once we recognise a specific situation we have engage in many times before, we respond in a ritualistic fashion (Goffman, 1959), before central cognitive processing even has a chance to kick in. The power of such norms does absolutely rule behaviour - the prospect of a large personal gain may induce a generally law-abiding trustee to defect on a particular transaction, as may conflicting

priorities - e.g. the need for money to help someone very important). The type of norms and degree to which they are adhered to will differ between people from different groups - if this is not understood by both parties, this will lead to mis-matched expectations between trustors and trustees, which is likely to disrupt the trust-building process (Vasalou & Riegelsberger, 2008)

While norms generally foster trustworthy behaviour among actors within a group, they are also a potential source of vulnerability. Social engineering and phishing attacks exploit social norms - such as obeying authority, helping a person in distress, and reciprocity (trust someone who has indicated they trust you). Violating ingrained norms can have a high psychological cost for users, which is why many users feel powerless to resit such attacks. A chat message from someone who appears to be a colleague you owe a favor to, who accidentally locked himself out of an internal system, and just needs your help in accessing some confidential information appeals to a number of internalised norms - and the psychological cost of not helping in that situation is quite high. Often, attacks exploiting social norms combine this with other factors to make them even more effective: exploiting established habits, and diverting users' attention (see Principle 1).

How should systems designers deal with the social norms that users bring to the systems they design? In our view, you ignore them at your peril, since together with user goals (Principle 1), they drive user behaviour. And security mechanisms that require over-riding of social norms are always in danger of being ignored - telling people not to help each other is not really an option. A first step has to be awareness of what values and internalised norms are likely to drive user behaviour in a particular interaction, and check whether these may present a source of vulnerability. Once the vulnerabilities have been identified, the mechanisms that require behaviour not in line with relevant internalised norms have to be explicitly brought to users' attention (Flechais et al. 2005). All participants in such transactions have to recognize the situation as 'special', i.e. as one in which they adopt a different role and follow a specific protocol that gives them no 'wiggle room'. Such explicit protocols - *"these are the rules everyone has to abide, no exceptions"* help to isolate the required behaviour from the psychological cost of breaking social norms. To achieve this, the situation as experienced by the user must beclearly distinguished from the every-day situations in which the internalised norms apply. This can be achieved through specific credentials that have to be exchanged, changes in user interfaces or physical tools, changes in titles or role names. Airline pilot communications around safety-critical maneuvers illustrate this point well: both actors wear uniforms, are identified by their roles rather than names, and the language is very precise and highly pre-scripted. Double-checking on the other pilot's actions is not a sign of doubt or lack of trust, but part of the role being enacted.

7. Users may not act rationally - but that's no excuse for not keeping them safe

Dhamija et al. 2006 found that participants commonly ignore trust indicators and warning signals on websites. We need to consider that the trustor's decision in situations of uncertainty is framed by their perceptions of risk and potential gain. In a recent study (Brostoff et al. 2010), we found that users may even ignore essentially

effective warning indicators, as long as the potential gain appears immediate and sufficiently high. A user who covets Tiffany earrings, but cannot afford them, will be very hard to deter from visiting a site that advertises Tiffany earrings at 75% discount. And as long as the site is plausible in terms of trust indicators (professionally designed to mimick Tiffany branding, sporting a logo and photos of items the user reognises). Once the wishful thinking combines with goal-driven behaviour, trustors will need a very specific and authoritative indicator that the earrings she is about to buy will not really be Tiffany.

The faulty trust decisions in the situations above are part of a much broader phenomenon that has been demonstrated across a wide set of contexts (Thaler & Sunstein, 2008). Given the opportunity, many people will opt for an immediate gain, even if it incurs a potentially large, but future and vague loss. Often-cited examples are smoking or drinking: the benefits are immediate, the costs are in the distant future and also uncertain - consequently many people choose to smoke and drink to such a degree that jeopardises their future health. The finding can also be translated to areas more typically associated with rational decision-making and planning: pension schemes. Sunstein & Thaler have shown that people put off the inconvenience of updating pension plan arrangements, even though such negligence can have huge impact on their future spending power and welfare.

We have illustrated in the examples above that the same effects apply when it comes to guarding oneself against phishing attacks or safeguarding one's privacy. Trustors do not really want to give their email, address, date of birth, and social security number to an online retailer, but if that's what it takes to get something they really want, they are likely to throw caution to the wind.

Another weakness that is commonly exploited in such attacks is people's sentimentality about sunken costs: Once a trustor has invested money, time and effort in in a transaction, they may be reluctant to accept that the best thing to do is cut their losses. Crime investigators are regularly stunned to fund that (often elderly) people keep sending money to scammers that tell them they have won something, and they'll get it if only they send this amount, then a bit more, then a bit more. Other examples include online shops that advertise goods at low prices, but won't deliver unless the buyer agrees to buy additional accessories or warranties at hugely inflated rates at the same time. Since the reluctance to write off losses is a strong motivator that is often exploited by dishonest trustees, and designers should consider whether in some transactions, if the trustor does not fulfill, it would be best to block or mediate any futher contact between trustor and trustees.

What conclusions can we draw for the design of systems? First, there are many ways of influencing trustors' perceptions of risk and benefit, and currently dishonest trustees have pretty free range in exploiting them. Not all of the exploits are criminal, as in the case of legitimate organisations who employ similar techniques in their e-commerce checkout or product sign-up processes - flagging additional charges late in the transaction, requiring just a few more bits of personal information before confirming a booking or purchase.

These observations show that there is further need for contextual properties in online transactions, in the form of regulation or self-regulation: future costs and dangers should be disclosed, a sense of urgency should not be

artificially created, users need to have an opportunity to review their own decisions and amend them at a later stage. Many of these principles are already part of codes of conduct in many industries: in the UK, for example, distance selling regulations give buyers a 10-day period during which they can cancel online purchases.

8. Why current trust signals are often useless

While system designers would do good taking in account people's common failings in decision-making when designing a system, to protect them from the worst consequences, as outlined above, there is also the inverse problem of security system designers accusing users of acting irrationally, when - in fact - on closer inspection they are doing exactly what helps them achieve their goal.

A good example for this situation are the frequent unspecific warnings that something may - or may not be - wrong with a website (such as "invalid certificate"). More often than not, these warnings are triggered by self-signed or expired certificates, and are thus not a reliable indicator of a dishonest trustor. When strong warnings are found to be unreliable, they lose their ability to function as trust signals - trustors will stop paying attention to them. The plethora of current alerts thrown at users through pop-up boxes, combined with a high false positive rate, has totally eroded their value as trust signals - rather, it has conditioned users to just 'swat' those alert boxes away without processing them. From a user perspective this is rational behaviour (Herley 2009): most often nothing bad will happen (or if it does it's hard to causally link it to the alert) and quickly getting rid of the alert will allow users to get on with their main task. Of course such reactions, when trained on a sufficient number of useless alert boxes can become habitual (see Principle1), so that even meaningful or serious alerts get clicked away without being read.

What's more, the warning messages are often phrased in a way that cannot be meaningful for users in their current decision-context. For a warning to be effective, trustors need to be told *"This site is not registered as an official ticket agency, and thus the likelihood you will get any tickets is zilch"*. Of course, this requires a rather more sophisticated infrastructure than is currently available. But we feel that is what is needed if we want genuinely trustworthy online transactions. Current attempts to deter users with unspecific, but strongly worded warnings that have a high rate of false positives are counterproductive.

9. Technology shifts trust signals, user perceptions lag behind

One practical implication of the distinction between symbols and symptoms (see Section 3.2) is that new technologies may shift symptoms (i.e. existing reliable indicators of trustworthiness) to become mere symbols (i.e. indicators that have significance only based on convention). An example of this is use of the padlock icon: a physical padlock can be touched, weighed and pulled by a person to assess its robustness. It is thus a more reliable indicator of the degree to which something is secured, whereas a collection of pixels that visually

resembles a padlock is a mere symbol of something being secured - the user has no way of assessing the level of protection offered. As stated in 3.2, symbols can be useful - but only if their abuse is prevented by institutional safeguards. Even if that is the case - which given the global nature of the Internet, and the difficulty of cross-border enforcement, is not always effective - the link between the symbol and the security it provides is far more indirect. For trustor, assessing the relationship between signifier and signified is much more work, and thus may not happen (see Principle 1).

More generally, the shift from symptoms to symbols caused by new technologies has implication for system designers. The risks that new technologies and services bring will be evaluated by most trustors using in the same terms they used to evaluate the transaction before introduction of technology. Adams and Sasse (1999) found that, after the introduction of small cameras and networked video transmission in offices and at conferences regularly caught out people who assumed they were in a private situation, when in fact others were watching them. It is not clear to many trustors that information shared once in an online world maybe available to anyone thereafter, forever. Many trustors have outdated notions of practical obscurity, which govern their decisions, such as: *"Who would be interested in finding something out about me online? They would never find information about me among the millions of people on the web"*.

The physical world also still guides many users' assumptions of location and identity. Visually inspecting a URL to see whether a link directed you to the site you expected to go to is no longer a reliable way of establishing the identity of a destination since Internationalized Domain Names (IDN) were launched by ICANN in 2008. What visually looks like google.com may in fact be a combination of western and other alphabets and point to a phishing site. Trustors may treat the URL as a symptom of identity - something that technically linked in such a way that it cannot be altered - when, in fact, they are looking at something that only visually, or symbolically, resembles the identity they were seeking to confirm.

The implications for designers is that they need to understand the assumptions that users will draw on when they make decisions about new forms of interaction, and clearly flag differences between potentially outdated interpretative models users may be applying, and the way a new tool or service actually works. Designers often use metaphors that users are familiar with as a way to make new systems easier to understand; but from a trust perspective, they may be doing them a disfavour since it re-inforces existing assumptions about trust signals and system behaviour, rather than making them aware of the way in which the new technology shifts these.

10. Reputation systems are no panacea for fostering trustworthy behaviour

One of the first applications of trust-warranting properties in online environments has been the development of reputation systems. They operates on the assumption of stable identities, and allow members of a community to publicly accumulate accounts of past interactions with specific trustors and trustees. These past accounts are often expressed in reputation scores, star ratings, and sometimes include free-text information. As these systems reduce the cost of collecting, distributing, and inspecting reputation information, they can support trust-building

among users who have not interacted before, and who have no plans for future interactions (i.e. no temporal embedding).

However, reputation systems also have their well-documented shortcomings. Sellers (trustees) on eBay, for instance, can build up high reputation scores through a large number of well-executed, but financially negligible transactions, and then cash in on this high trust score by defecting in a high value final transaction before leaving the system, and possibly returning later under a new identity (see Principle 9; Dellarocas & Reskin, 2003). eBay have continually tried to improve their system, and now allow trustors to inspect not only reputation scores, but also the transactions that the score is based on. This, however, can be subverted through shill transactions, where traders appear to buy from each other and leave positive feedback, even though in the real world, not money or goods are exchanged.

However, reputation systems are imperfect ways of providing stable identities (see Principle 9). eBay, for example, uses credit cards as a verification mechanism, but malicious trustees can register credit cards at different addresses, and thus leave the system after defecting on a transaction, and the re-enter back with a new and 'clean' identity.

The social mechanisms that rely on the 'wisdom of crowds' by using behaviour in the system alone can be easily subverted by attackers (Stajano & Wilson 2009). Reputations can be hijacked - indivudual users' accounts may be hijacked to present a credible trustor, and even companies. There have been cases of established web businesses with good reputation being taken over by attackers, who reduce prices to very attactive level, take orders and payments, but never deliver.

The conclusion is that over-reliance on any indicator of intrinsic properties - such as reputation - puts it at risk, because it becomes just too valuable, and just a target for attack. To avoid tis, designers should support a number of different ways of signalling intrinsic properties, and also support different groups and networks of users in building their own feedback in a less structured way.

References

Adams, J. 1995.Risk. UCL Press.

Adams, A. and Sasse, M. A. 1999. Taming the Wolf in Sheep's Clothing: Privacy in multimedia communications. Proceedings of ACM Multimedia'99, Orlando, Florida November, 1999, pp. 101 - 107. ACM.

Axelrod, R. 1980. More Effective Choice in the Prisoner's Dilemma. Journal of Conflict Resolution 24(3), 379-403.

Bacharach, M. and Gambetta, D. 1997. Trust in Signs. Working Paper. University of Oxford.

Bacharach, M. and Gambetta, D. 2003. Trust in Signs. In: Cook, K. S. Trust in Society. Russell Sage: New York, NY,

Berg, J., Dickhaut, J., and McKabe, K. 2003. Trust, Reciprocity, and Social History. Games and Economic Behaviour 10, 122-142.

Brostoff, S., Jennett, C. and Sasse, M. A. 2010: An evaluation study of the SOLID anti-phishing tool. Technical Report UCL Department of Computer Science, UK.

Chernoff, H., 1973, The Use of Faces to Represent Points in K-Dimensional Space Graphically, *J*ournal of the American Statistical Association 68 (342): 361–36

Corritore, C. L., Kracher, B., and Wiedenbeck, S. 2003. On-line trust: concepts, evolving themes, a model. International Journal of Human Computer Studies 58(6), 737-758.

Dellarocas, C. & Reskin, P., 2003, 1st Interdisciplinary Symposium on Reputation Mechanisms in Online Communities (April 2003), Cambridge, MA, USA.Deutsch, M. 1958. Trust and suspicion. Journal of Conflict Resolution 2(3), 265-279.

Dhamija, R., Tygar, D. and Hearst, M. 2006. Why Phishing Works. Proceedings of CHI 2006

Fehr, E., Fischbacher, U., 2003,(The nature of human altruism: Proximate patterns and evolutionary origins, Nature

Flechais, I., Riegelsberger, J., Sasse, M. A., 2005, Divide and Conquer: The role of trust and assurance in the design of secure socio-technical systems *Proceedings of the 2005 workshop on New security paradigms, p*p. 33 - 41

Friedman, J. W. 1977. Oligopoly and the Theory of Games. North-Holland Publishers: Amsterdam.

Friedman, B., Kahn, P. H., and Howe, D. C. (2000). Trust Online. Communications of the ACM 43(12), 34-40.

Fukuyama, F. 1999. Social Capital and the Civil Society. In: 2nd Conference on Second Generation Reforms. IMF: Washington, DC, US.

Giddens, A. 1990. The consequences of modernity. Stanford University Press: Stanford.

Goffman, E, 1959, The presentation of self in everyday life, Doubleday

Granovetter, M. S. 1973. The Strength of Weak Ties. American Journal of Sociology 78, 1360-1380.

Granovetter, M. S. 1985. Economic Action and social Structure: The Problem of Embeddedness. American Journal of Sociology 91, 481-510.

Herley, C. 2009. So Long, And No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. Proceedings of NSPW 2009.

Kumaraguru, P., Cranshaw J., Acquisti, A., Cranor, L.F, Hong, J., Blair M. A.and Pham, T. 2009, School of Phish: A Real-World Evaluation of Anti-Phishing Training. Proceedings of SOUPS 2009.

Luhmann, N. 1979. Trust and Power. Wiley: Chichester.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. 1995. An Integrative Model of Organizational Trust. Academy of Management Review 20(3), 709-734.

McKnight, D. H. and Chervany, N. L. 2000. What is Trust? A Conceptual Analysis and An Interdisciplinary Model. In: American Conference on Information Systems. 827-833.

Ratnasingam, P. and Pavlou, P. A. 2004. Technology Trust in Internet-Based Interorganizational Electronic Commerce. Journal of Electronic Commerce in Organizations 1(1), 17-41.

Raub, W. and Weesie, J. 2000a. The Management of Durable Relations. Thela Thesis: Amsterdam.

Raub, W. and Weesie, J. 2000b. The Management of Matches: A Research Program on Solidarity in Durable Social Relations. Netherland's Journal of Social Sciences 36, 71-88.

Riegelsberger, J., Sasse, M. A., and McCarthy, J. 2003b. The Researcher's Dilemma: Evaluating Trust in Computer-Mediated Communication. International Journal of Human Computer Studies 58(6), 759-781.

Riegelsberger, J., Sasse, M. A., McCarthy, J. D., Trust in Mediated Interactions, The Oxford Handbook of Internet Psychology, 2007, pp. 53-69.

Riegelsberger, J., Sasse, M. A., McCarthy, J. D., Do People Trust Their Eyes More Than Ears? Media Bias in Detecting Cues of Expertise, *Extended Abstracts CHI'05,* 2005.

Stajano, F. & Wilson, P. (2009): Understanding scam victims: seven principles for systems security. University of Cambridge Technical Report UCAM-CL-TR-754 ISSN 1476-2986

Thaler, R. H, and Sunstein, C. R., 2008, Nudge:improving decisions about health, wealth, and happiness, Yale University Press

Vasalou, A., Riegelsberger, J., 2008, Recovering trust and avoiding escalation: an overlooked design goal of social systems, *Conference on Human Factors in Computing Systems - CHI*, 2008, pp. 3333-3338.