Divide and Conquer: The role of trust and assurance in the design of secure socio-technical systems

Ivan Flechais
Oxford University Computing Laboratory
Wolfson Building
Parks Road
UK – Oxford OX1 3QD
+44 (0)1865 285302
ivan.flechais@comlab.ox.ac.uk

Jens Riegelsberger
Department of Computer Science
University College London
Gower Street
UK – London WC1E 6BT
+44 (0)20 7679 0351
j.riegelsberger@cs.ucl.ac.uk

M. Angela Sasse
Department of Computer Science
University College London
Gower Street
UK – London WC1E 6BT
+44 (0)20 7679 7212
a.sasse@cs.ucl.ac.uk

ABSTRACT

In order to be effective, secure systems need to be both correct (i.e. effective when used as intended) and dependable (i.e. actually being used as intended). Given that most secure systems involve people, a strategy for achieving dependable security must address both people and technology. Current research in Human-Computer Interactions in Security (HCISec) aims to increase dependability of the human element by reducing mistakes (e.g. through better user interfaces to security tools). We argue that a successful strategy also needs to consider the impact of social interaction on security, and in this respect trust is a central concept. We compare the understanding of trust in secure systems with the more differentiated models of trust in social science research. The security definition of "trust" turns out to map onto strategies that would be correctly described as "assurance" in the more differentiated model. We argue that distinguishing between trust and assurance yields a wider range of strategies for ensuring dependability of the human element in a secure socio-technical system. Furthermore, correctly placed trust can also benefit an organisation's culture and performance. We conclude by presenting design principles to help security designers decide "when to trust" and "when to assure", and give examples of how both strategies would be implemented in practice.

1. INTRODUCTION

The aim of security is to identify risks, and devise countermeasures that effectively mitigate the risks to the assets of a system. Security countermeasures are traditionally distinguished into avoidance, deterrence, prevention, detection, reaction and insurance.

To counter threats effectively, however, any countermeasure has to *function correctly* and *be dependable*. We define the two properties as follows:

- *Correctness*: the designed countermeasures will neutralise the threat if working as intended.
- *Dependability*: the degree to which designed countermeasures are working as intended.

"A computer is secure if you can depend on it and its software to behave as you expect" [10]. Although this definition is open to debate (because it implies that security exists in the reader's expectations of computer and software behaviour), it is useful to highlight the importance of dependability in computer security. It has been argued that an emerging sentiment in security research is *"correctness" is not the issue; "dependability" is'* [5]. The point is that the ability to know how the system is going to behave is now being recognised as very important, in addition to building a system that actually counters threats.

A secure system is part of a wider socio-technical system whose goal is the achievement of a production task [1, 6, 25, 29]. A socio-technical system has both human and technical components working together to achieve production tasks, as well as achieving the enabling task of securing that system effectively [6][1]. Dependability is therefore determined by the degree to which this socio-technical system behaves in the way it is expected to. Technical components are designed, and their behaviour is easier to predict then that of the human element. (Though technical systems created by putting together several sub-systems can exhibit unexpected emergent behaviours.) However, the effectiveness of social engineering attacks [17], and reports of people's failure to comply with organisational security policies, demonstrate that the behaviour of the social element of a secure system is currently much less dependable.

New Security Paradigms Workshop '05, September 20-23, 2005, Lake Arrowhead, California, USA.

[1] In systems theory, security would be classed as one of the supporting measures designed to ensure the long-term survival of the system [7].

It is often assumed that one way of increasing the dependability of the security in a system is to limit the role that people have in that secure system, using technical countermeasures to replace the human element because they are seen to be more dependable. Governments, for instance, currently deploy biometric systems because they are better at detecting an individual presenting a passport that is not theirs. However, replacing people by technology is not feasible, nor is it always desirable: people may function less reliably on some tasks, but they are also more flexible, and often perform multiple functions. Thus, removing the human element may weaken the overall security. [27], for instance, describes how a US border control officer identified a would-be terrorist – who had his own passport and was not under suspicion – because she detected abnormal behaviour *"something about him just wasn't right"*. Table 1 shows that there are technical and social countermeasures for every dimension of security (prevention, detection, reaction and deterrence). While the number of technical solutions is increasing, they cannot be expected to fully replace social countermeasures.

Technical countermeasures, if selected and configured correctly, perform well on repetitive security tasks, such as access control, virus checking or integrity checking. They become less effective in less well-defined security tasks, such as anomaly detection (i.e. intrusion detection), and detecting hitherto unknown undesirable events, or their pre-cursors. People, although lacking the accuracy and being prone to fatigue and boredom, can be very flexible and extremely effective – especially since deterrence tends to be a social mechanism (for example prosecution and jail). A design strategy for a secure socio-technical system must be to use the strengths of both components, and avoid their weaknesses.

	Category	Description	Example
Technical countermeasure	Prevention	Stop attacks from happening	Firewalls, access control, etc.
	Detection	Notice and identify an attack	Intrusion detection systems, Automatic terrorist profiling
	Reaction	Stop or mitigate an attack	Automated response mechanisms linked to intrusion detection systems
	Deterrence	Discourage abuse	The visibility of technical countermeasures. E.g. CCTV
Social countermeasure	Prevention	Stop attacks from happening	Don't share passwords, lock your screen, have security guards on the gate
	Detection	Notice and identify an attack	Sysadmins, alert users, audit checking
	Reaction	Stop or mitigate an attack	Sysadmins or emergency response teams
	Deterrence	Discourage abuse	Prosecution, financial ruin, job loss, prison

Table 1. Technical and Social countermeasures

Another point to consider is that people in a socio-technical system are not static components – people evolve, and interact to form social subsystems. Security designers may consider the capacity to evolve a negative characteristic, since it means the behaviour of the human element is not consistent – e.g. an employee who has complied with security policies for many years can suddenly "turn bad" if his circumstances change or and exceptional temptation arises. At the same time, since most organisations operate in a constantly changing environment, the human ability to evolve is a necessary condition for their survival.

The interactions between people in any system are governed by *social norms* – rules by which people behave – and based on *values* – people's beliefs, for instance, about what is right and what is wrong. Norms can evolve in social systems over time, or can be designed. Law and security policies are examples of designed norms that govern behaviour. Norms that are not formalised, but are pervasive (common to most social systems) have evolved and are widely adhered to because, over time, they have turned out to be of advantage for the long-term survival of the system.

A prime example of such a norm is *trust.* In social sciences and economics, trust has been researched extensively over the past 4 decades. The resulting, widely accepted definition of trust is *"an attitude of positive expectation that one's vulnerabilities will not be exploited"* [4, 21]. The idea of "willingness to be vulnerable", as opposed to deploying a countermeasure, may initially appear to be anathema in the context of security. But it is worth noting that – in terms of well-established thinking in systems theory and social sciences – such a norm would not have evolved unless it was beneficial to the long-term survival of systems. There is ample evidence of the economic benefits of trust: high-trust systems are much more expensive to operate than low-trust ones [11]. This economic benefit is largely due to two factors:

1) Devising and operating countermeasures for every vulnerability is expensive. If two parties trust each other and neither party breaks the trust, they both lower their cost of interaction.
2) Trust is a pre-condition for the creation of *social capital* in systems. Social capital means that people in an organisation have shared values and a shared sense of responsibility for the well-being of a system, which reduces selfish behaviour and carelessness.

Given that security is there to ensure long-term survival of a system, the potential benefits of trust as defined in social science and economics are intriguing.

In security, on the other hand, trust is currently defined as a *"system or component […] whose failure can break the security policy"* [3]. This definition sees trust as a characteristic of a component, whereas in social sciences, it is a property of the system that forms as a result of interaction between agents of that

system. The security definition also implies that failure of such a component must be avoided, i.e. any risk to the dependability of the component must be mitigated. Human trust, however, means *not* mitigating the dependability risk that arises from the vulnerability. The security perspective is one of *assurance* – rather than trusting that the vulnerability will not be exploited. A "trusted system" is traditionally seen as one that is well secured, meaning that all known vulnerabilities have been removed or counteracted through security measures.

We argue in the remainder of the paper that both assurance and trust have a role to play in the design of a secure socio-technical system. Firstly, using both strategies, each in its rightful place, improves the economics of a secure system – trusting is cheaper, as long as that trust is well-placed, i.e. not exploited too often. Secondly, trust can actually improve the dependability of the human component, because it fosters the development of shared values and responsibility, and increases vigilance and the motivation to comply. Thirdly, a more detailed understanding of assurance and trust can help to design more effective assurance mechanisms, because the reduction of trust that usually accompanies heavy-duty assurance mechanisms can be counteracted through organisational design measures. Finally, this perspective gives designers a richer understanding of the vulnerabilities that affect the dependability of the human element of a socio-technical system, and puts a wider range of countermeasures at their disposal.

In the next section, we present a brief overview of some of the current research in the Human-Computer Interactions in Security (HCISec) community. We then present a detailed account of how trust relationships are built between actors (whether human or not), and what key factors foster well-placed trust. We will then present a discussion as to how trust and security factors can work together in order to make people more dependable in their application of security policies, and therefore less likely to become unwitting victims of social engineering attacks. We conclude by calling for more research into this area and introduce a number of design principles that may favour well-placed trust between organisations and their employees, as well as fostering a trusting environment within the organisation.

2. PEOPLE AND SECURITY

Kahn [13], cited by Anderson [2], *"attributes the Russian disasters of World War 1 to the fact that their soldiers found the more sophisticated army cipher systems too hard to use, and reverted to using simple systems which the Germans could solve without great difficulty"*. This statement seems to expound the notion that good security is hard to use.

Bruce Schneier [26], however, makes the point that *"security is only as good as its weakest link, and people are the weakest link in the chain"*, indicating that good security has to acknowledge the weaknesses of people. Other authors [1, 12, 17, 18, 30] also argue that secure systems are broken through human issues, such as bad security configuration. They state that ease of use is necessary in order to get people to behave securely, and that good security is not necessarily hard to use [1].

Consequently, the whole field of HCISec is largely focussed on building better tools [8, 28] and improving the user interfaces to these tools [12, 30]. This will undoubtedly improve the usability of security tools, and in turn improve security. However, we also believe that improving the user interface is only one of many changes designers have to make to improve the dependability of the human element in secure socio-technical systems.

When Saltzer and Schroeder in 1975 identified the need for *'psychological acceptability'* [23] in secure systems, they were referring to the need for better interfaces. However, *psychological acceptability* extends beyond user interfaces because a secure system is acceptable if the *user cost* (i.e. the sum total of the psychological, cognitive and physical load required of a user in a given task) of using it is not excessive compared to the *user benefits* (i.e. the incentives and advantages of engaging in a given task). This goes beyond the security user interface, and affects the user in the wider context of system use.

As mentioned above, in most organisations security is a secondary, *enabling task*, ensuring the continuity of the primary production task [24]. One of the costs of security is how much it will interfere with production tasks. Possible benefits of applying security might be avoiding penalties, or peer acceptance into a particular "security conscious" group. In organisations that prioritise productivity whenever there is a conflict with security, and which do not penalise those who break security policies, or reward those that do comply, the cost of applying security for an individual is high compared with the benefits. Unsurprisingly, people involved in the security of such organisations are less likely to behave as intended.

3. TRUST

What role does trust play in improving the dependability of the human element in socio-technical systems? The term 'trust' is frequently used in the security literature – for example when referring to trusted paths and trust chains. In contrast to this, as stated in the introduction, social science research defines trust *"an attitude of positive expectation that one's vulnerabilities will not be exploited."* [15, 16, 22]

A useful starting point when looking at the role trust plays in security is to identify which factors influence an actor's decision to engage in a trust relationship. (Actors can be people, but also organisations, institutions, and job roles – such as bank clerks, couriers or policemen). A trust relationship is only required when risk and uncertainty are present, i.e. when actors stand to lose something. At the same time, the trustor often expects to realise a gain if the transaction is successful. These factors can be observed in eCommerce transactions, where the customer pays the vendor in the expectation that the vendor will send the desired goods – which are often available at a lower price than from traditional retailers or unavailable locally. The customer cannot ensure compliance from the vendor and has to trust that they will keep their side of the bargain. The vendor, on the other hand, has the option to default on sending the goods and a number of factors can influence this decision.

Figure 1 from the research presented in [21], illustrates the factors which determine the mechanics of trust between a *trustor* (i.e. trusting actor) and a *trustee* (i.e. trusted actor). It consists of a number of factors that affect how trust is signalled, how these signals are understood and how they affect a given trust relation. The main factors consist of:

- Intrinsic properties
  - *Motivation*, *Ability*, *Internalised Norms* and *Benevolence*
- Contextual properties
  - *Temporal*, *Social* and *Institutional embeddedness*.

Both contextual and intrinsic properties play a role in the establishment of a trust relationship. Intrinsic properties refer to factors that are internal to the trustor and trustee, such as the propensity to take risks, the benefits of engaging in a trust relationship and the personal cost of breaking trust. Contextual properties refer to factors that exist outside both actors, such as law enforcement, expectations of future interactions or reputation.

A further important distinction to introduce at this stage is that between trust and *reliance*. Trust governs the early exchanges between a specific trustor and trustee. With repeated successful exchanges, the trusting stance – i.e. where the trustor is conscious of his vulnerability – is replaced by an expectation – or *reliance* – that the trustee will behave in a trustworthy manner. That is to say the trustor does not consider himself as vulnerable any more. The distinction is important because attacks on the human element in secure systems that exploit reliance differ from those that exploit trust.

3.1 Intrinsic Properties

3.1.1 Trustor: Motivation

Motivation refers to an actor's incentive for engaging in a trust relationship. It is affected by factors such as propensity to trust, perception of risk, benefits of engaging in the relationship and the availability of other options that may achieve similar results. These are subjective characteristics that vary betweens actors. Propensity to trust relates to the trustor's inclination to be trusting – some people are more inclined to be trusting, for instance for fear of offending the trustee by not doing so. [17] present many examples of social engineering attacks which exploit this.

The perception of the risk of engaging in a trust relationship refers to the potential for loss – not only financial, but also for example the psychological cost of having been naïve, or having been duped by attackers. The propensity for risk, again, differs from person to person, and some people break security policies simply because they enjoy taking risks; interestingly, most people are less likely to take risks on behalf of others [29].

Benefits capture what the actor stands to gain from a successful trust relationship, such as financial profit, a reduction in cognitive effort, time saving, etc. In security, there are examples of people disclosing passwords in exchange for a reward such as

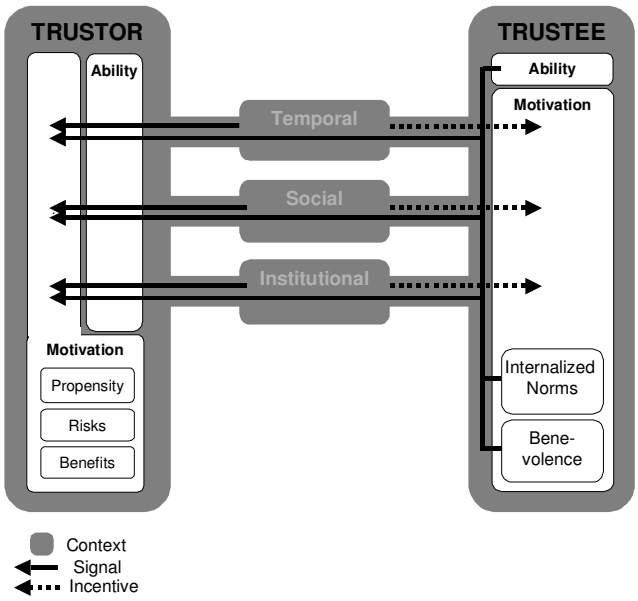

Figure 1. Model of Trust

chocolate bar[2], or giving access to their computer to a person that offers to fix a purported problem [17]. Finally, a critical factor as to whether actors engage in a trust relationship is whether these properties can be detected – i.e. if individuals perceive there is no benefit, they will not be motivated.

3.1.2 Trustor: Ability

For the trustor, this refers to the individual knowledge and understanding of the signals and situations that affect the formation of a trust relationship. For example, a trustor's assessment of the risk inherent in a given trust relationship is affected by past experience as well as new knowledge. An employee may be happy leaving their PC screen unlocked when a maintenance person is in the office, but may change this behaviour if subsequently customer data was found to have been downloaded from that PC.

3.1.3 Trustee: Ability

Ability refers to the trustee's ability to actually achieve a given task. The trustee may be willing but unable to actually perform in the manner expected. For instance, many people may be willing to keep separate passwords for different systems, until they find they are unable to recall them when necessary, and then resort to breaking security policies [25]. Unrealistic expectations of people's *ability* to behave in a dependable manner can also reduce dependability in a wider sense: security policies that require the impossible create resentment and lower people's general willingness to comply with security policies, i.e. it reduces their *motivation* to be dependable [1].

3.1.4 Trustee: Motivation: Internalised Norms

Actors have been observed to behave in a trustworthy manner despite not having any external incentive to do so. Partly this can be put down to habit, but also to *internalised norms* that affect

[2] http://news.bbc.co.uk/1/hi/technology/3639679.stm

that actor. These norms can induce the actor to behave in an untrustworthy manner, for selfish actors motivated only by immediate gain. They can also induce the actor to behave in a trustworthy manner, a trait that can be referred to as *integrity*:*"… the trustee adheres to a set of principles that the trustor finds acceptable."* [15]. For example, some people based on their upbringing, are more diligent in following the rules.

The most immediate application of this understanding to security is that most people do not break a trust relationship lightly. If employees interpret being given access to a system as being trusted to look after an organisational asset, their internalised norms can create a barrier to breaking that trust. Security policies that do not explain countermeasures in terms of protection of assets, but try to assure behaviours by threatening sanctions, on the other hand, do not have this effect (see also section 3.3).

People act on internalised norms based, for instance, on their upbringing. This is important to consider for two reasons (1) because people are likely to violate security policies that demand behaviour that conflicts with their internalised norms, and (2) because some attacks play on internalised norms. [29] provides an interesting example of the first category: a company's security policy stated that employees must lock their screens whenever leaving their desks. In small offices shared by 2-3 people, locking your screen whenever you left your desk was interpreted by co-workers as a sign that they were not trusted. Rather then jeopardise relationships with their co-workers, employees preferred to break the security policy.

When security policies conflict with internalised norms, security designers need to manage the conflict. In the financial sector, for instance, there are many examples where individual employees are not trusted, e.g. no single employee can open the safe. If company employees understand that this policy is necessary to comply with external regulations, or to protect the reputation of the organisation, it de-personalises the fact that employees are not trusted to open the safe, making it clear that the lack of trust is *"business not personal"*.

[17] provides examples of attacks that exploit internalised norms, either by pretending to be a co-worker in need of help, or by someone in authority, and who are therefore to be trusted. Many email scams and phishing attacks use the same approach. A good strategy for increasing dependability of employees in the face of such attacks is to institute simple, reliable rules for mutual authentication, and a supportive point of contact for no-fault reporting and clarifying rules.

Many professions, such as medicine, instil norms in their practitioners and organisations should consider promoting a set of norms that supports their security goals. An example of such a norm is "our customers entrust their data and privacy to us, and we have a shared responsibility to protect them at all times". Consistent promotion of such norms can lead to them taking priority over other internalised norms, at least in the organisational context (see section 3.3).

3.1.5 Trustee: Motivation: Benevolence

"Human behaviour in romantic relationships is an example of trustworthy action motivated by strong feelings of benevolence. In such relationships the well-being of the other forms part of one's own gratification. Benevolence – albeit to a lesser degree – also applies to relationships between work colleagues or friends" [21]. In relationships of *benevolence*, actors do not expect immediate or equal returns, but this sentiment only evolves over time, and after a number of successful trust exchanges. This factor can be crucial in breaking security policy, [17] presents several examples of how social engineering attackers groom their targets by appearing to be benevolent – e.g. selflessly helping to fix a problem on the target's PC (which, of course, the attacker has created in the first place), and exploiting the resulting trust relationship. Social engineering relies on the target's *benevolence* and willingness to be helpful in order to break policies.

3.2 Contextual Properties

3.2.1 Temporal embeddedness

This is the notion that two actors' decision to engage in a trust relationship is affected by their expectation of future interactions. This is one of the reasons why, for example, many disgruntled employees are willing to vandalise and cause damage to systems they have access to, since they have no expectations of future interactions with their employers. Some organisations have "exit protocols" that make sure that people who are leaving the organisation cannot exploit trust that was extended to them as employees. In most organisations, however, there is no systematic checking that all access to systems has been removed, for instance, that any shared passwords have been changed. Similarly, people are often not made aware that meeting former colleagues in a social context can lead to disclosure of sensitive information.

3.2.2 Social embeddedness

Social embeddedness represents the group interactions and ensuing reputation that an actor gains from his behaviour towards members of that group. The incentive to behave in a trustworthy manner is no longer linked to future interactions with a single actor, but to future interactions with actors likely to hear of their reputation. Within organisations, this can be a powerful motivator for a newcomer to conform to the existing security behaviours (or lack thereof) inside their immediate peer group. [29] reports that newcomers' behaviour with respect to security policies invariably follows that of members of their immediate work team, even when they have undergone security training as part of their induction. The desire to "fit in" the immediate work environment is usually stronger. This emphasises the need for security awareness and training to be given continuously to all employees, as opposed to just giving it to newcomers.

3.2.3 Institutional embeddedness

This refers to the organisations (e.g. employee's company, ethics committees, or consumer rights groups) or institutions (e.g. law) that have the power to sanction untrustworthy behaviour or behaviour that is below expectations. Here the significant factor for engaging in a trust relationship is governed by the type, strictness and severity of punishment. This type of sanctioning is governed by strict rules defining what the institutions' expectations are in a given situation. An example of this could be the threat of being excluded from a professional group as a result of objectionable behaviour under a given code of practice.

This type of deterrence is currently the most widely used means of *assuring* compliance to the security policy. However, [25] points out that this is currently often ineffective because those in executive positions fail to comply with security policies. High-level managers often feel their time is too valuable to comply with 'petty' security regulations. The effect is that other employees will not interpret breaking of security policies as a breach of the trust that the organisation places in them. Futhermore, imposing sanction on some members of an organisation, but not on others, prevents the development of a shared set of values that could foster a better security culture, and thus increase dependability.

3.3 Trust vs. Reliance

In trust exchanges, trustors do not expect their vulnerabilities to be exploited, but they are usually aware that they are taking a risk, and balance this against the expected benefits. After several successful exchanges, however, trustors develop an expectation that the trustee is reliable. The distinction is important because once a trustor comes to rely rather than trust, the awareness of risk is lowered. Some attacks on the human element in secure systems exploit this by inducing reliance through repeated trust exchanges, and attacking once reliance is established. One example are attacks on reputation systems of internet auction sites, when dishonest traders build a positive reputation through a series of low-value exchanges, and then default on subsequent higher-value ones. Another strategy is to impersonate a trustee on whom the trustor has come to rely. Phishing attacks exploit reliance by setting up web sites that look similar to ones that the target is familiar with, using similar URLs or symbols. The oldest attack to harvest login credentials replicated the login screen, and skimming attacks on cash dispensers also exploit the reliance customers have placed on these machines.

Security designers need to be aware that certain familiar cues will induce reliance and trigger habitual responses in people. If organisations employ such clues, they need to be difficult to fake. Security awareness campaigns can sensitise people, and training can improve their ability to recognise such attacks. With the increasing sophistication of attacks, however, mutual authentication is likely to become the only effective countermeasure for preventing people from falling prey to impostors.

3.4 Assurance vs. Trust

Assurance consists of the *contextual* factors that organisations can put in place to ensure a specific outcome to a trust exchange. These are currently mainly restricted to detecting and sanctioning undesirable behaviour. *Trust* on the other hand is based on an understanding of the *intrinsic* properties that pertain to a given actor. In high-security environments, organisations seek to establish whether an actor is intrinsically trustworthy by conducting background checks. These are intended to determine whether the actor has any past evidence of law-breaking, indicative of individuals whose integrity may be less than satisfactory. These investigations focus on past behaviour; the discussion in section 3.1.4 highlights that the analysis of an individual's internalised norms may give some clues as to how likely they are to break trust, and thus to their dependability.

As stated by [21], organisations are more productive if they have *social capital* [19] – i.e. trust that is based on shared informal norms that promote cooperation [9]. Some authors claim that reported failures of systems to yield the expected productivity gains in organisations [14] partially stems from a reduction in opportunities to build social capital [20].

Currently, security policies are generally designed to encourage actions that can be readily interpreted as untrusting. For example, refusing to share a password with a colleague, locking your computer screen or checking the credentials of a technician are all signs of distrust in any usual setting and are considered to be basic security practices. As discussed in section 3.1.4, it is important to de-personalise this interpretation and replace it with an understanding of organisational assets and security requirements.

The design of current secure systems rarely considers the need for – or existence of – trust between the different operators running the system. The attitude that prevails in system design is that the operator of the system must and will perform a task, with little to no thought going into how this will affect him in the wider context of his organisation. Ignoring this issue can damage the formation of social capital in the organisation, or even provide a means of forming social capital through the breaking of security practices, i.e. employees bonding together in the knowledge that everyone is breaking the rules with them.

For example, it may be that for confidentiality purposes a medical data provider has specified a policy that separates different kinds of medical data (i.e. general health, sexual health, aids, cancer, etc.) and restricts access to these. There are cases where a particular organisation has a number of different projects utilising these resources, all needing different access privileges so as not to be given unnecessary information. This can put that particular organisation in a position where the different projects combined have access to all areas of information from the provider, yet the policy is designed to prevent some projects from accessing parts of this information. So, although the organisation as a whole is trusted with the totality of the confidential medical information, the policy requirements result in particular individuals being allowed more access than others. Lacking any justification for this measure, this can be interpreted as a lack of trust of an organisation in its employees, as well as resulting in increasing the administration costs of maintaining the access control mechanism.

4. ANALYSING TRUST AND SECURITY

4.1 Where Trust and Security meet

The role of security in an organisation is to dependably handle the threats to the assets of that organisation. In order to do this, both technical and social countermeasures are necessary, and ensuring that these countermeasures are actually applied is of equal importance. Many organisations hold the misconception that security is best achieved if there is no need to trust any of the employees within the organisation, because the rules and procedures in place would be sufficiently reliable as to avoid any risk of employees acting undesirably.

Well-defined, repetitive and predictable tasks, lend themselves well to creating and enforcing a policy that compels employees to

apply security, whilst preventing them from abusing the system. A good example of this can be seen in the banking sector which has evolved a vast number of procedures, both technical and social, to prevent employees from stealing money. The disadvantage of using these types of *contextual* measures is that it takes away a lot of flexibility, makes organisations slow to respond to new situations, and is a costly means of operating.

In areas where job requirements are vague, or there is a specific need for flexibility, these kinds of rigid policies cannot be made to work because they are too complex, constraining or expensive. In these cases, the only available option is to choose, encourage and trust employees to behave in a secure manner, rather than enforce it. This can also be complemented by monitoring in order to detect whether employees are actually complying with the policy.

4.2 Breaking Trust

As we have seen in section 3, security policies can require people to behave in a manner that is bound to be interpreted as not trusting others. For example, a trustor requires a trustee to divulge his password in order to allow the trustor to finish some urgent work. In this exchange the *motivation* for the trustor to engage in this trust relationship is that he has a high potential *benefit* – i.e. finishing the urgent work – and the other *options* are more inconvenient and time-consuming – getting his password reset or reissued.

The trustee has the *ability* to divulge his password, and may feel *benevolent* which may influence him to choose to divulge in order to help a colleague. On the other hand the trustee may have a degree of *integrity* that prevents him from behaving in such a way as to disobey the security policy. The *motivation* to refuse or accept to share the password is also affected by external factors.

Expectations of *future involvement* may tip the balance in favour of breaking the security, since it is very likely that the trustee will interact again with the trustor as they are colleagues. The trustor may also be a part of a larger group of colleagues and in cases where security is not important to this s*ocial group*, they might give the trustee a bad reputation or affect the relationships between him and the group should he decide to refuse. The final factor is the degree to which the *organisation* detects and punishes transgressions and rewards good behaviour.

Should the trustee refuse to violate the security policy, unless the trustor understands and agrees with the *motivation* to do so, he may feel untrusted and untrustworthy, which can create tensions between the two actors, and will definitely hinder the creation of

Design principle	Description	Relevant Property
Simplifying security	Make the task of behaving securely easier through better tools and simpler interfaces but also through simple policy rules – exceptions to the rules can be sources of confusion and abuse.	*Ability* *Motivation* (other options)
Promoting a security culture	A security culture should be encouraged by ensuring that the security policy is neither excessive (i.e. for every countermeasure there is a corresponding threat) nor unfair (i.e. the boss is allowed to avoid security measures). In addition to this monitoring and checks should be made regularly to ensure the policy is in use, transgressions are detected and punished according to a published code of conduct, and secure behaviour is rewarded.	*Ability* *Social embeddedness* *Organisational embeddedness* *Motivation* (avoiding punishment, benefiting from reward)
Participative Security	In situations where a decision has to be made as to what security countermeasures to adopt, involving the relevant stakeholders in the decision making process may improve the feeling of trust from the organisation and the *motivation* to apply the resulting countermeasures.	*Ability* (Improved knowledge of security) *Social* and *Institutional embeddedness* *Motivation*
Group membership Group identity	Specifically grouping people into security groups, together with their own responsibilities and rewards can make security a more immediate concern for employees. By making the groups smaller expectations of future interactions are greatly increased, thus harnessing that particular factor.	*Ability* *Temporal*, *Social* and *Institutional embeddedness* *Benevolence*
Educating employees about security	By providing employees with training as to what is expected and required and what are the threats.	*Ability* *Motivation* (Perception of Risk) *Benevolence*

Table 2. Principles for fostering dependable behaviour from the social elements of a secure system

social capital. In this case three trust signals can positively influence the adherence to the security policy without harming the trust relationship between both actors:

- Providing an alternative to initiating the trust exchange in the first place. This can be done by giving the trustor an easy way of accessing the systems he needs, for example by reissuing his password in a timely manner or providing a limited access based on a temporary password and monitoring the activity of the trustor. This is the kind of approach that HCISec is trying to achieve by making it easier to use secure systems.
- Having a security conscious culture within the peer group. Both actors can relate to this, even if they do not overtly recognise it (i.e. if everyone is careful with their passwords and refuses to divulge them, then the peer pressure to behave in the same way is significant).
- Ensuring that the detection and punishment for breaking the security rules are effective. Very stringent enforcement of security policies will result in adherence to the policy. This is a very straightforward means of preventing rule breaking because it is easily understood by both trustor and trustee, who have a lot more to lose than gain.

In banking, the stringent security measures in place do not create tensions amongst staff because it is well understood that the detection and punishment for a transgression is taken very seriously. This in turn can foster an environment where no one breaks the rules, thereby reinforcing the *motivation* to avoid transgression. Stringent enforcement can only happen in areas where the expectations are as well defined as the punishments. As we have seen above, the disadvantage of this approach is that it stifles flexibility and this makes it inappropriate for a significant number of jobs that require security.

4.3 Middle Ground

As seen in section 3, there are two extremes of security:

- Assurance: complete control over what employees must and can't do, together with stringent enforcement.
- Trust: no control over what employees can do, and only trust and encouragement for them to behave in a secure manner.

The problems start to occur when trying to secure a system that exists in the middle ground of being able to support well-defined security policies, whilst still requiring a degree of flexibility. In cases like this, where the security policy in place is either not well-defined (in order to maintain flexibility), it is essential that the enforcement of that policy be both strictly specified and applied to everyone in the organisation. In addition, it is essential to foster an environment which encourages employees to behave in a trustworthy manner. Table 2 describes a set of design principles that make use of the trust warranting factors identified in section 3.

Following our presentation and analysis of the factors influencing trust in secure systems, we believe further research in this area would undoubtedly yield greater insights into secure socio-technical system design.

5. CONCLUSION

Getting a secure system to behave dependably is a complex task. Assurance mechanisms can achieve a degree of success, but in most real-world situations, organisations either cannot afford the costs of maintaining such a stringent system, or need to be flexible. This means that these systems have to rely on people behaving in a secure manner. We have looked at the field of trust and identified a number of factors that affect an individual's propensity to behave in a trustworthy manner. We are convinced that these factors can be applied to improving the dependability of an individual's security behaviour, and have presented a number of design principles aimed at addressing this.

6. REFERENCES

[1] Adams, A. & Sasse, M. A. *Users Are Not The Enemy.* Communications of the ACM 1999. Vol. 42, No. 12 December

[2] Anderson, R. *Why Cryptosystems Fail.* ACM Conf.Computer and Communication Security CCS'93 1993. pp 215-227.

[3] Anderson, R. *Security Engineering: A Guide to Building Dependable Distributed Systems*. 2001. Wiley.

[4] Bacharach, M. & Gambetta, D. *Trust as Type Detection.* C.Castelfranchi & Y.Tan (Eds.), Trust and Deception in Virtual Societies 2001. pp 1-26. Dordrecht: Kluwer.

[5] Baker, D. *Fortresses built upon sand.* Proceedings of the New Security Paradigms Workshop 1996.

[6] Brostoff, S. & Sasse, M. A. *Safe and Sound: a safety-critical approach to security design.* New Security Paradigms Workshop 2001.

[7] Checkland, P. & Scholes, J. *Soft Systems Methodology in Action.* 1999. John Wiley and Sons Ltd.

[8] Dourish, P. & Redmiles, D. *An Approach to Usability Security Based on Event Monitoring and Visualization.* Proc.New Security Paradigms Workshop (Virginia Beach, VA) 2002.

[9] Fukuyama, F. *Social Capital and the Civil Society.* 2nd Conference on Second Generation Reforms 1999. Washington, DC: IMF.

[10] Garfinkel, S. & Spafford, G. *Practical UNIX and Internet Security*. 1996. O'Reilly.

[11] Handy, C. *Trust and the Virtual Organization.* Harvard Business Review 73(3) 1995. pp 40-50.

[12] Ka-Ping, Y. *User Interaction Design for Secure Systems.* 2002. http://zesty.ca/sid

[13] Kahn, D. *The Codebreakers*. 1967. Macmillan.

[14] Landauer, T. K. *The Trouble with Computers: Usefulness, Usability, and Productivity*. 1996. Cambridge, MA: MIT Press.

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

[16] McAllister, D. J. *Affect- and Cognitation-based Trust as Foundations for Interpersonal Cooperation in Organizations.* Academy of Management Journal 1995. 38(1) , pp 24-59.

[17] Mitnick, K. D. & Simon, W. L. *The Art of Deception: Controlling the Human Element of Security*. 2003. John Wiley & Sons Inc.

[18] Poulsen, K. *Mitnick to lawmakers: People, phones and weakest links.* 2000. http://www.politechbot.com/p-00969.html

[19] Putnam, R. D. *Bowling Alone: The Collapse and Revival of American Community*. 2000. New York: Simon & Schuster.

[20] Resnick, P. *Beyond Bowling Together: SocioTechnical Capital.* HCI in the New Millenium 2002. pp 242-272. Boston, MA: Addison-Wesley.

[21] Riegelsberger, J., Sasse, M. A., & McCarthy, J. *The Mechanics of Trust: A Framework for Research and Design.* International Journal of Human Computer Studies 2004. 62(3) , pp 381-422.

[22] Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. *Not so different after all: A cross-discipline view of trust.* Academy of Management Review 1998. 23(3) , pp 393-404.

[23] Saltzer, J. H. & Schroeder, M. D. *The protection of information in computer systems.* IEEE 1975.

[24] Sasse, M. A. *Computer Security: Anatomy of a Usability Disaster, and a Plan for Recovery.* CHI 2003 2003.

[25] Sasse, M. A., Brostoff, S., & Weirich, D. *Transforming the 'weakest link': a human-computer interaction approach to usable and effective security.* BT Technical Journal 2001. 19 , pp 122-131.

[26] Schneier, B. *Secrets and Lies*. 2000. John Wiley & Sons.

[27] Schneier, B. *Beyond Fear Thinking Sensibly about Security in an Uncertain World*. 2003. Copernicus Books.

[28] Smetters, D. K. & Grinter, R. E. *Moving from the design of usable security technologies to the design of useful secure applications.* New Security Paradigms Workshop.September 23-26, 2002, Virginia Beach, VA 2002.

[29] Weirich, D. & Sasse, M. A. *Pretty Good Persuasion: A first step towards effective password security in the real world.* New Security Paradigms Workshop 2001.

[30] Whitten, A. & Tygar, J. D. *Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0.* Proceedings of the 8th USENIX Security Symposium, August 1999, Washington 1999.