Trust: A Collision of Paradigms

L. Jean Camp
Kennedy School of Government
Harvard University
Cambridge, MA
jean_camp@harvard.edu

Helen Nissenbaum
University Center for Human Values
Princeton University
Princeton, NJ
helen@princeton.edu

Cathleen McGrath
College of Business Administration
Loyola Marymount University
Los Angeles, CA
cmcgrath@lmu.edu

Abstract

The technological challenges of securing networks are great, as recently witnessed in widespread denial of service and virus attacks. The human reaction to these attacks may be either a loss of trust or a willingness to tolerate increasing risk having weathered one assault. Examining human and computer interaction with a focus on evaluations the human response to loss of trust is a key part of the search for more secure networks. The success of current efforts to design appropriate security mechanisms depends as much on an understanding of human extensions of trust to computers as it does on an understanding of underlying mathematics. However, the former has not been sufficiently examined.

In this work we survey the findings in social psychology and philosophy with respect to trust. We introduce three hypotheses that remain unanswered with respect to the manner in which humans react to computers. We discuss potential design revisions in light of findings from other disciplines. Then we conclude by noting that research which empowers users in order to be their own security manager may be based on a fundamentally flawed view of human-computer interaction. We close by encouraging designers of computer security systems to examine the humans, which these systems are intended to empower, and recommend that any security system be built on the basis of understanding of human trust provided by the social sciences.

Introduction

Although there has been progress in the quest to build more secure and trustworthy systems, regular news of intrusions, breaches, and rogue attacks serve as reminders that there is a great deal more to be done. Experts focus on the considerable technological challenges of securing networks, designing strategies, building mechanisms, and devising policies. Although these efforts are essential, we here suggest that trust and security would be even better served if designs more systematically addressed the (sometimes irrational) people and institutions who are critical components of networked information systems. Accordingly, efforts at securing these systems should involve not only attention to machines, networks, protocols, and policies, but also a systematic understanding of how the social agents (individuals and institutions) participate in and contribute to the security and trust of networks.

This is not to imply that technical work in security ignores the role of people and institutions in networks and network security. Rather that in addition to the technical, good network security requires a more systematic account of the ways people feature into network security than it has previously incorporated. The goal of our paper is to offer a way in which to begin to address the ubiquity of human engineering by understanding how current security systems may be built on hypotheses of human action which are not sustainable. Certainly this has been recognized with respect to the fact that humans are unreliable sources of random information.

We examine the study of trust from social and philosophical perspectives. This leads to identification of implicit assumptions about the ways people behave, trust, and conceptualize security. We show that these assumptions conflict with results and arguments found in theoretical and empirical work in philosophy and social science.

The Variable of Trust

We develop three hypotheses where technology and social science seem to be on a collision course. However, each of these hypotheses at its core points to a common point of collision: technologists often assume that humans are attentive, discerning, and ever-learning. Philosophy argues that humans are simplifiers, and this implies that humans will use trust of machines to simplify an ever more complex world. Social science argues that humans may slowly lower barriers against trust, rather than refining them.

To be specific, theories relating social capital and trust predict that, if computers are perceived as elements of a single undifferentiated network, then trust in computers will increase as computing experience increases. If these theories of

human behavior are applicable to computer/human interaction then computer security mechanisms must be built with the assumption that individuals will be too likely to trust untrustworthy machines, and that this risk-taking behavior will increase over time.

Conversely in computer science there has been an implicit assumption that humans learn to manage their own network security as individuals. If humans do learn to differentiate between servers then increased experience on the network will correlate with a greater ability to distinguish trustworthy and untrustworthy machines. Mechanisms from content selection (e.g. PICS), and privacy calculations (e.g. P3P), to public key systems (e.g. PGP) require that humans learn to manage trust on a machine-by-machine or transaction-by-transaction basis.

If the view of the social sciences is correct then the autonomy provided to users in an end-to-end network may in fact undermine the autonomy of a naive user, rather than enhance it, by exposing the user to risk which the naive user cannot reasonably be expected to manage. A user who cannot secure his or her machine from malicious code and malevolent crackers cannot be said to be autonomous.

Research has found that interface design (e.g. Kiesler, Sproull & Waters, 1996), group affiliation (e.g. Dawes, McTavish & Shaklee, 1977) and communication (e.g. Kerr & Kaufman-Gilliland, 1994) influence the extension of trust. While these studies focus on the effect of computer mediation on the extension of trust, they do not address the issue of the trustworthiness of the underlying computer technology with which individuals interact. The rapid advance of computer performance and connectivity has meant that individuals are often interacting with and depending on more computer systems, with a greater diversity in computer hardware and computer software. In addition the owners of these machines are increasingly diverse as the Internet is adopted for business across the globe and across the demographic range of industrialized nations.

As computer systems become more integral to individual action, social interaction, and commerce, the study of trust must extend to understand how individuals extend trust to computers and computer systems. Since the early work on computer mediated trust, human/computer interaction has become extremely common. Bloom (1998) proposes that the human willingness to expose information to a computer will usher in a new age of social science, in which data accuracy is ever increasing, as computers become ubiquitous. As computer use becomes more widespread and computer users more sophisticated, human willingness to divulge information may suggest an overall increase in users' trusting behavior regarding computer mediated interaction. Such observed behavior may indicate a decreased ability to distinguish between various machines and thus suggests that computer security policies and mechanisms which require active learning on the part of users may prove woefully inadequate.

In contrast, other research suggests people now have large and increasing concerns with privacy and security in information technology (e.g. Wacker, 1995; Walden, 1995; Hoffman & Clark 1991; Compaine, 1988; Computer Science and Telecommunications Board, 1994). Examinations of computer systems show that security protections are inadequate (Office of Technology Assessment, 1985; Office of Technology Assessment, 1986; National Research Council, 1996). Professionals in computer science, law and business (e.g. Wacker, 1995; Walden, 1995; Anderson, Johnson, Gotterbarn & Perrolle, 1993; United States Council for International Business, 1993) point to privacy and security as the stumbling blocks of electronic commerce.

Privacy and security concerns reflect a growing unwillingness to expose information to computers, and suggest greater discernment on the part of users called on to increasing trust the machines which increasingly dominate transactions in daily life. The emergence of electronic commerce and public key-based encryption systems increases the need for computer security and appropriate evaluation of computer trustworthiness. Public key cryptography uses certificates to link (usually) a person, an electronic key and some attributes. With public key infrastructures individual computer users are expected to become security managers. The design decision is based on the assumption that users are increasingly discerning of distinct machines. Yet this core assumption remains unexamined, as technologies which require users to select which individual public keys, key hierarchies and computer systems to trust. In designing public key infrastructures for the mass market, it is critical to understand the direction and nature of individual user's approach to computers with respect to trust. The implications of previous studies suggest that beliefs commonly implemented in computer security systems should perhaps be reversed -- for example, the interface should be purposefully less attractive to avoid lulling users into potentially inappropriate high trust behavior. Understanding how the trust that computers engender will evolve over time is critical to the appropriate evolution of security mechanisms.

The Internet as Self-Organizing

In order to argue that social theory results should have a significant impact on the design of security systems we consider the definition of trust provided by social theory. Axelrod (1984) poses the question, "Under what conditions will cooperation emerge in a world of egoists without central authority?" (p.3). His results suggest that the willingness to extend trust initially and to display forgiveness at some point after a defection are important to the maintenance of a cooperative social group. We argue that the Internet illustrates trust as exhibited by the self-organization of egoists who choose to extend trust in order to connect. The Internet has central authority with respect to the assignment of domain names and Internet protocol addresses. However, there is no central authority to govern the daily interactions on the Internet.

We argue that the emergence of connectivity and ordered communication illustrates the applicability of the studies of social theory to the Internet and in particular to the design of security systems. On at least three levels, trust is necessary and extant on the Internet. First, at the nuts and bolts level of the router system, users must explicitly and implicitly trust that each link of the underlying technology of the Internet will behave in the way it is expected to behave. Second, users must trust that other people will behave in ways that uphold the community norms in the absence of central authority to enforce norms. Finally, users must trust that institutions - such as Internet businesses - will conduct themselves in ways that are conducive to productive ongoing transactions. Trusting the nuts and bolts level, means trusting the underlying infrastructure of the Internet, which in turn is made up of the collaborative effort of several computers connected together. A router that must be maintained by individuals at the site controls each top-level domain.

Any single router can seriously impede the functioning of the Internet. (For example, a router on the East Coast once decided it was in Berkeley and became a black hole for what would otherwise have been smoothly-flowing traffic.) Changing a single parameter in a single router table can cause significant damage to network traffic. At the level of infrastructure, the integrity of the Internet can be compromised by individual error or guile. That it is rarely so compromised underscores the high level of collaboration and mutual trust underlying the Internet's functioning. The TCP SYN flooding attack was an open secret on the Internet for many years. Anyone who knew TCP could have implemented the attack, but for more than a decade no one did.

The picture is similarly challenging in the case of trusting the people of the Internet. For example, every USENET newsgroup is self-governing and therefore vulnerable to the bad behavior of a small subset of users. Every group member must adhere to the rules of participating in the USENET newsgroup. For example, members may only post on topics relevant to the group, and they must treat others in the group with respect. Periodically, the ground rules of group participation are posted to the entire group, but no mechanism for enforcement of the rules is in place. Some USENET newsgroups have been disbanded, and others have descended into eternal flamewars or spam pits because users did not adhere to the rules. Yet many continue to flourish.

As necessary as trust is in Internet commerce as it is implemented today, trusting virtual institutions poses special challenges as well. The consumer has no way to validate the existence of a business, nor the comfort offered by the location and presentation of a storefront. Items can not be examined in a tactile manner before purchase. The existence of the item may be pure fiction, and transmitting one's credit card number to such a merchant is indeed an exercise in trust.

A question of particular interest to the design of secure systems in how people individuate the agents with which they interact. How users individuate networked machines is relevant to when and how they extend trust, in the first place, and how this trust is affected by betrayals of various kinds. In other words, do people extend trust to computers as single agents, or do they individuate and distinguish among them? Alternatively do people consider all computers elements of a single network with out distinguishing between what are, in fact, very distinct machines? Outside of the question of the way users experience their interactions with computers, system designers may reasonably think that users should distinguish among individual computers as they do among individual people, or individual institutions. This is because, despite attempts at quality control and reliability, computers differ from each other from the moment they are shipped from the factory floor. They differ in terms of operating systems, exposure to the environment, exposure to viruses, and history of use.

All of this makes it likely that individuals will have different experiences with different computers. By "surfing the

net" users are choosing to interact with many different computers. And of course, the trustworthiness of the people themselves behind the computers on the Internet covers the range of humanity. However, it may be the case that individuals do not differentiate among the different computers or different human agents behind the computer with which they interface, knowingly or unknowingly. If users do not make distinctions among different computers that they use, then they may extend trust or refuse to extend trust using past information and experiences that are not entirely applicable to the new situation.

Defining Trust from the Social Science Perspective

The social sciences offer us a definition of trust which may be useful in computer security; and is certainly useful for this discussion. Coleman's (1990) definition of trust accounts for the rational action of individuals in social situations. Coleman's definition of trust has four components:

1. Placement of trust allows actions that otherwise are not possible.

2. If the person in whom trust is placed (trustee) is trustworthy, then the trustor will be better of than if he or she had not trusted. Conversely, if the trustee is not trustworthy, then the trustor will be worse off than if he or she had not trusted.

3. Trust is an action that involves the voluntary placement of resources (physical, financial, intellectual, or temporal) at the disposal of the trustee with no real commitment from the trustee.

4. A time lag exists between the extension of trust and the result of the trusting behavior.

Coleman's definition is consistent with a rational decision making model. His is a behavioral definition rather than an affective decision. In this framework, trust is not a feeling as much as an action. If a person would be no worse off after placing resources in the hand of the trustee and having the trustee cheat, then trust is not an issue. So, for example, trust would not be an issue if an individual delivers a message to a client and also asks a colleague to deliver the same message to the client. In this case, the individual does not have to trust his or her colleague because the message has already been delivered, and no bad consequence will occur if the colleague does not hold up his or her end of the agreement.

Notice that “trusted” in the social sciences has exactly the same meaning of “trusted” in computer science. That is, something that is trusted is trusted exactly because if it fails there is a loss. Except in the case of computer security there is often an assumption that the trusted third party is trustworthy, and there is no such assumption in social theory.

Often the costs of safeguarding against untrustworthy behavior are so high that the only solution is to extend trust to others. This is currently the case in routing, USENET newsgroups, and commerce described above. Trustees must make judgments about whether or not the people with whom they enter agreements are likely to uphold them.

Trustees may not hold up their end of the agreement because they lack the ability to take the agreed-upon actions, committing error. Alternatively, trustees may not hold up their end of the agreement because they have made a decision to defect on the agreement, or cheat, to improve their own welfare at the cost of the trustee, acting with guile.

Error and guile are two possible causes of the breakdown of trust agreements. For example, in the case of Internet routers, an individual at a single site, may be not able to handle the volume of traffic going through the domain and make an error in routing. On the other hand, an individual at the site may decide to take down another site for self-interested reasons. The result is the same in both cases. However, the individual extending trust may react differently in response to error versus guile. We propose that reactions to computer “betrayals” as opposed to “betrayals” by obvious human action will result in different forgiveness behavior. People's decisions to trust computers may be affected by their perceptions of the difference between computers and humans in error making and acting with guile. It is a commonly held belief that computers only replicate human error and that computers can be easily monitored to find the source of error. Also, computers are not perceived to act with guile.

Previous research has supported the hypothesis that people are more trusting of computers than of other people. For example, people disclose more information and more accurate information during interviews with computers than

during interviews with humans (Sproull and Kiesler, 1991). However, these studies do not consider people's willingness to distinguish between trustworthy and untrustworthy computers (or, reliable and unreliable computers) in the same way that people are willing to characterize different individuals as either trustworthy or untrustworthy. Thus there has been a disconnect between the critical trust questions in computer security and those questions as framed in other fields.

Besides the connections we wish to make between the proposed study and past empirical studies of trust in computer mediated environments, our work is informed by social-theoretic and philosophical work on trust. Social theorists, like Niklas Luhmann (1979) stress the trial-and-error nature of the development of trust, suggesting that starting with a baseline desire (indeed need to trust) people begin with a readiness to trust. This initial readiness to trust is then put to the test in transactions with others, where it is either confirmed or undermined by their experiences with the particular object of their trust. Another relevant thesis that emerges out of both philosophical and social scientific work on trust is that trust is not as vulnerable to incompetence as it is to bad intention. That is, people are ready to forgive harms they may have suffered due to incompetence far more quickly and readily than harms they perceive to have been caused by the bad intentions of others. (See for example, Becker 1996 and Slovic 1993.)

Conflicting Assumptions

Implicit and unexamined assumptions about trust are embodied in many widely prominent technical security techniques and mechanisms. Yet, work in philosophy and social science on trust offers reasons for thinking that at least some of these assumptions are wrong. In this section we discuss three hypotheses, showing how they have informed existing security mechanisms and policies and suggesting ways that the mechanisms and policies might be altered in light of this knowledge. The cases fit our paper's theme, namely that optimal security systems would draw on what is known about trust in non-technical literatures and paradigms.

> **Hypothesis I:** In terms of trust and forgiveness in the context of computer-mediated activities, there is no significant systematic difference in people's reactions to betrayals that originate from human actions, on the one hand, and computer failure, on the other.

According to this hypothesis, in terms of effects on trust in computers and computer-mediated activity and readiness to forgive and move on, people do not discriminate on the basis of the origins of harms such as memory damage, denial of service, leakage of confidential information, etc. In particular, it does not matter whether the harms are believed by users to be the result of technical failure, on the one hand, or human (or institutional) malevolence.

For example, key revokation policies and software patches all have an assumption of uniform technical failure. Consider key revokation. A key may be revoked because of a flawed initial presentation of the attribute, a change in the state of an attribute, or a technical failure. Currently key revokation lists are monolithic documents where the responsibility is upon the key recipient to check. Often, the key revokation lists only the date of revokation and the key. The social sciences would argue that the three cases listed above would be very different and would be treated differently. Consideration of that possibility leads to a key revokation system which may better fit human consideration of trust, and manage risk more effectively as well.

Consider the case of an incorrect initial attribute. In this case, the possibility of malevolent action is most likely. Consider identity theft, since identity is a favored attribute linked to public keys (and was in fact required by the first X.509 standard). Identity theft would call for more than revokation at the date of discovery. In a web of trust system; for example, the revokation should be able to be broadcast or narrowcast to anyone whose key or reputation is authentication by the stolen identity. Any extension of cumulative trust enabled by the use of the key should be removed, and this should occur recursively until the entire result of the stolen identity is removed. Alternatively any accounts set up or configured with this key should be terminated. The capacity to create additional accounts and thus implement a domino of trust extensions is exactly the feature which makes identity theft attractive. Thus, key revokation schemes should take into account this capacity when evaluating methods for addressing the revokation of a particular key.

Consider a change in the state of an attribute. For example, a particular employee may be unauthorized to charge a particular account after a sudden, unexpected, or particularly unpleasant termination. In this case, again, accounts that

may have been created for the duration of the certification should be reconfigured. An example may be an account at B2B exchange that requires certification at account initiation and considers the key lifetime, as set by the employer, as the appropriate duration of a valid account. Noting that this is a sub-optimal policy by the exchange is not likely to prevent flawed policies from being adopted; in particular when the interest of the businesses and the exchange is to accept risk in order to prevent denial of service. Recall that the Electronic Funds Transfer Act was initiated by exactly this type of change in attribute and malevolence, although in that case the malevolence resulted from divorce and not employment termination. The card issuer had a policy that expected individuals to know in advance how long the attributes -- in that case the marriage -- would last. In contrast, given a technical failure of a lost key all that would be necessary is that future assertions by the holder of the subverted key. By having a single standard key revokation systems implement the assumption that there is no significant systematic difference in people's reactions to betrayals which originate from human actions, on the one hand, and computer failure, on the other.

With respect to software patches, the possibility of a purposefully malevolent alteration of the code is not considered. The social sciences would argue that such cases require a different level of active response and oversight than technical error made in the market equivalent of good faith. For example, a bug purposefully placed by hackers who had access to Microsoft's source code would presumably be meant for harm; while the other 63,000 bugs in Win2k (Foley, 2000) could be considered either minor or less likely to enable malevolent action. Thus a discovery of a malevolent bug should result in active contact with all customers who had installed the product and technical support to enable effective patching; while the standard policy of customer-driven seeking and downloading could be adequate for other cases.

The hypothesis makes sense from a purely technical standpoint. Certainly good computer security should protect users from harms no matter what their sources and failure to do so is bad in either case. Yet a closer examination, based on an understanding of social theory, yields a more complex problem space and more nuanced solution to the problem of key revokation or patch distribution.

Nevertheless, there are good reasons for questioning the hypothesis. One is related to a view that a number of researchers hold about trust: that it should be reserved for the case of people only, that people can only trust (or not trust) other people not inanimate objects.

These researchers suggest that we use a term like confidence or reliance to denote the analogous attitude people may hold toward things like computers and networks. To the extent that this is more than merely a dispute over word-usage, we are sympathetic to the proposal that there are important differences in the ways trust and confidence or reliance operate (See, for example, Seligman; Nissenbaum; and Friedman, Kahn and Howe.) One reason to reserve the concept of trust for a relation between people is the role motives and intentions seems to play in it. Various works on the subject of trust have discussed this. The philosopher Lawrence Becker, for example, has argued that far more relevant to our readiness to trust are the motives and intentions we perceive others to have than actions and outcomes. So, for example, if we believe that things have gone wrong as a result of incompetence, our trust will be far less affected than if we believe ill-will to be behind it. Psychologists Paul Slovic and Tom Tyler, in separate works, have demonstrated similar themes, namely, that the way people see intentions mediating outcomes is significant for trust and forgiveness. What this means for our purposes is that people's trust would likely be affected differentially by conditions that differ in the following ways: cases where things are believed to have gone wrong (security breaches) as a result of purely technical glitches, as compared with cases where failures are attributed to human engineering, as compared with cases where evil intention is seen as the cause of harm. Even within these categories, there is quite a range of difference. A number of the cases involving identifiable human agents (i.e. including incompetence and malevolence) can, for example, be seen as points along a continuum rather than as instances of entirely non-overlapping categories. To briefly illustrate, a security breach which is attributed to an engineering error might be judged accidental and forgiven if things went wrong despite considerable precautions taken. Where, however, the breach is due to error that was preventable, we might react to it in a way that is more similar to our reaction to malevolence. Readers familiar with categories of legal liability will note the parallel distinctions that the law draws between, for example, negligence versus recklessness.

Efforts at designing security mechanisms and policies which reflect the varieties of human judgments and reactions and be sensitive to these distinct conditions will be more robust in real-world environments. This hypothesis also has implications for the design of intrusion detection systems. It implicitly suggests that up to a point the false negatives in intrusion detection are more dangerous than false positives. Currently there are risks in intrusion detection systems of allowing activity which is suspicious versus the risk of producing too many false positives. If humans perceive much

malicious activity to be simple reliability failures, a higher level of suspicion generated by false positives would be preferable to a false sense of security, as undetected attacks will be unduly accepted and forgiven.

> **Hypothesis II:** When people interact with networked computers, they sensibly discriminate among distinct computers (hosts, websites), treating them as distinct individuals particularly in their readiness to extend trust and secure themselves from possible harms.

In terms of best practices for security, it makes most sense for people to view distinct remote computers as distinct individuals, each one warranting independent evaluation. Yet, there are several reasons that converge on a quite different story suggesting that users tend to view networked computers as constituting a more homogeneous system. Social theory predicts that individuals' initial willingness to trust and therefore convey information in the context of a web form will depend more on the characteristics of the individual and interface than the perceived locality of or technology underlying the web page. An empirical study of computer science students also demonstrated that experience with computers increases a willingness to expose information across the board.

What this means is that users, even those with considerable knowledge and experience, tend to generalize broadly from their experiences. Thus, positive experiences with a computer generalize to the networked system (to computers) as a whole and presumably the same would be true of negative experiences. In other words, users draw inductive inferences to the whole system, across computers, and not simply to the particular system with which they experienced the positive transaction. Such a finding would have grave implications for the design of user-centered security. Security systems which empower the user, for example, ActiveX, to select between trustworthy and untrustworthy code may not, in fact, be empowering if humans do not differentiate between machines. Human centered security mechanisms may prove to offer no more autonomy to the naive user than the option to perform brain surgery at home would offer medical autonomy to the naive patient. In fact, the argument that alterable code is not empowering to the user has been presented in the case of applications (Clark and Blumenthal, 2000). This tendency to generalize across computers has other implications for security strategies. It suggests that we should be thinking of ways to impress users with the distinctiveness of different machines so that they realize that trustworthiness of one is independent of trustworthiness of another.

In particular the Secure Sockets Layer and the pop-up windows as implemented in all currently and previously predominant browsers encourage users to consider the network to consist of two elements: secure and trustworthy pages versus insecure and untrustworthy pages. This is done by providing a uniform graphic to display at every site with no customization for user or site. The combination of the "lock" in the lower right-hand side and the notice of “leaving a secure” page encourages users to view all sites which use SSL to be equivalent in terms of trustworthiness.

In a related issue, that of ensuring that the person at one end of a connection is indeed connected to the host as believed, a useful solution for this problem has been proposed. Tygar and Whitten (1996) propose window personalization to prevent proxy attacks or Java Trojan Horses from stealing passwords. A similar window personalization could require that the installation of SSL includes a selection of a JPEG image to be included as part of the 'lock' image. This would communicate to the user that no two SSL-using sites are, in fact, the same. Furthermore, the deletion of this image would identify any redirections, for example from a conference site to a secure payment mechanism site -- a transition that now appears seamless to the SSL user. An examination of social theory suggests that the implementation of a program which has only increased security in the near term (SLL) could prove problematic in the long term by encouraging users to treat all machines with SLL as equally trustworthy.

> **Hypothesis III:** Over time and with experience users will tend toward greater discernment among distinct remote computers.

According to this hypothesis, the tendency to draw narrow inferences based on experience with remote computers will increase with users' level of experience with computers and computer mediated interactions. Computer experience alone cannot increase the tendency toward greater discernment among remote computers until the design of those parts of security mechanisms that users experience clearly signal differences among distinct computers. Current design encourages users to continue to generalize broadly on the basis of experience with individual cases. They simply will have more experience. This reduction is reinforced by theories of social capital addressing a broader social context. This work suggests that when decisions to trust individual members of a community are vindicated, these positive experiences will generalize to the community as a whole and thus will contribute to social capital.

It is often noted that if telephone systems still required operator assistance, the services of every man, woman, and child in the United States as operators would be required to support today's traffic on the public switched telephone network (PSTN). Similarly, the evolution of computers into ubiquity requires a decrease in the level of human labor as system operators. In the case of the PSTN, there were no requirements for alterations in human trust, as the smart network addressed issues of trust and security in billing and dialing. In fact, almost every man, woman and child in America is a telephone operator. Instead of requesting a location or number, we enter seven or more digits to enable a connection to the end user whom we seek.

In the packet-switched world end users must evolve into network operators. The switch from human-to-human requests (as dominates system operation today) to human-to-machine requests is far more problematic when the machine is multi-purpose. This is compounded by the requirement that humans become security managers. The capacity of humans as security managers is assumed to be high when hypothesis three is assumed correct. However, social theory and philosophy argue that hypothesis three is incorrect. If humans monotonically increase trust then user-managed security systems which monotonically increase trust are problematic.

Consider an implementation of cumulative trust, as with PGP or Lilith. In both cases the user begins with a small set of trusted parties and expands this set of trusted parties as these trusted parties vouch for others. In no cases are the system implemented with a requirement for a reset. That is, at each moment as trust accumulates it becomes more likely that the trust is being extended to an untrustworthy participant. A requirement that the machine effectively reset its trust barriers; for example, by requiring that the user select a predefined size for a set of initial trusted parties before the set is defined anew, could mitigate against the tendency of humans to increase trust for all computers. Assuming long term use, and the human tendency to be increasingly trusting, the social argument for a reset function is strong; although the technical argument is weak at best.

Consider the case of the Platform for Privacy Preferences. P3P allows a user to do business with a site which has privacy practices which follow the user's preferences. A natural result would be for a user to be informed, “To use this site you must enable privacy preference n,” just as today sites commonly recommend closed standards or lower security settings (e.g. accepting cookies) for interaction. Such a site may be one which has a particularly high draw. Eventually users may decrease their privacy thresholds so that P3P offers little or no protection. Again, a reset mechanism is called for. At the least, privacy settings should be lowered at a site-by-site basis when they are lowered, or lowered for a specific duration after initially being set. Conversely, if the hypothesis is correct then privacy protection increases should be implemented across all sites without a temporal limit.

Conclusions

In this work we have offered and supported three hypotheses with respect to the manner in which humans extend trust to computers. We have shown that the hypotheses are assumed to be correct in social theory and philosophy, and at least implicitly assumed correct in computer security implementations. For each hypothesis we have offered design suggestions which would align the computer science with the social science.

The first hypothesis was, “In terms of trust and forgiveness in the context of computer-mediated activities, there is no significant systematic difference in people's reactions to betrayals which originate from human actions, on the one hand, and computer failure, on the other.” In this case the hypothesis lead to criticisms of common key revokation practices. The second hypothesis was, “When people interact with networked computers, they sensibly discriminate among distinct computers (hosts, websites), treating them as distinct individuals particularly in their readiness to extend trust and secure themselves from possible harms.” This hypothesis lead to recommendations that visual identifiers which indicate that some particular mechanism is in use integrate signals which encourage users to differentiate between machines. The third hypothesis was, “Over time and with experience users will tend toward greater discernment among distinct remote computers.” This hypothesis is the most radical in terms of the differences between computer and social sciences. In general this hypothesis calls for caution in the implementation of user-managed computer security mechanisms. Specifically, this hypothesis, if correct, would argue that any user-managed system that tends to monotonically increasing trust would, over time, be completely subverted by user tendencies to extend trust.

In each case a social science and philosophical hypothesis had direct technical implications for the design of a purely technical system to implement trust. If it is not possible to design a computer security system without assumptions about human behavior then the design of computer security systems should be informed by philosophical and social science theories about trust.

References

Anderson, R. E., Johnson, D.G., Gotterbarn, D. and Perrolle, J., 1993, "Using the ACM Code of Ethics in Decision making," *Communications of the ACM*, Vol. 36, 98- 107.

Abric & Kahanês, 1972, "The effects of representations and behavior in experimental games", *European Journal of Social Psychology*, Vol 2, pp 129-144

Axelrod, R., 1994, *The Evolution of Cooperation*, HarperCollins, USA.

Becker, Lawrence C. "Trust in Non-cognitive Security about Motives." *Ethics* 107 (Oct. 1996): 43-61.

Blaze, M., Feigenbaum, J. and Lacy, J., 1996, "Decentralized Trust Management", *Proceedings of the IEEE Conference on Security and Privacy*, May.

Bloom, 1998, "Technology Experimentation, and the Quality of Survey Data", *Science*, Vol. 280, pp 847-848

Boston Consulting Group, 1997, *Summary of Market Survey Results prepared for eTRUST*, The Boston Consulting Group San Francisco, CA, March.

Clark & Blumenthal, "Rethinking the design of the Internet: The end to end arguments vs. the brave new world", *Telecommunications Policy Research Conference*, Washington DC, September 2000.

Coleman, J., 1990, *Foundations of Social Theory*, Belknap Press, Cambridge, MA.

Compaine B. J., 1988, *Issues in New Information Technology*, Ablex Publishing; Norwood, NJ.

Computer Science and Telecommunications Board, 1994, *Rights and Responsibilities of Participants in Networked Communities*, National Academy Press, Washington, D.C.

Keisler, Sproull & Waters, 1996, "A Prisoners Dilemma Experiments on Cooperation with People and Human-Like Computers", *Journal of Personality and Social Psychology*, Vol 70, pp 47-65

Dawes, McTavish & Shaklee, 1977, "Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation," *Journal of Personality and Social Psychology*, Vol 35, pp 1-11

B.Friedman, P.H. Kahn, Jr., and D.C. Howe, "Trust Online," *Communications of the ACM*, December 2000/Vol. 43, No.12 34-40.

Foley, 2000, "Can Micrsoft Squash 63,000 Bugs in Win2k?", ZDnet Eweek, on-line edition, 11 February 2000, available at http://www.zdnet.com/eweek/stories/general/0,11011,2436920,00.html.

Fukuyama F., 1996, *Trust: The Social Virtues and the Creation of Prosperity*, Free Press, NY, NY.

Garfinkle, 1994, *PGP: Pretty Good Privacy*, O'Reilly & Associates, Inc., Sebastopol, CA, pp. 235-236.

Hoffman, L. and Clark P., 1991, "Imminent policy considerations in the design and management of national and international computer networks," *IEEE Communications Magazine*, February, 68-74.

Kerr & Kaufman-Gilliland, 1994, "Communication, Commitment and cooperation in social dilemmas", *Journal of Personality and Social Psychology*, Vol 66, pp 513-529

Luhmann, Niklas. "Trust: A Mechanism For the Reduction of Social Complexity." *Trust and Power: Two works by Niklas Luhmann*. New York: John Wiley & Sons, 1979. 1-103.

National Research Council, 1996, *Cryptography's Role in Securing the Information Society*, National Academy Press, Washington, DC.

Nissenbaum, H. "Securing Trust Online: Wisdom or Oxymoron?" Forthcoming in *Boston University Law Review*

Office of Technology Assessment, 1985, *Electronic Surveillance and Civil Liberties* OTA-CIT-293, United States Government Printing Office; Gaithersburg, MA.

Office of Technology Assessment, 1986, *Management, Security and Congressional Oversight* OTA-CIT-297, United States Government Printing Office; Gaithersburg, MA.

Seligman, Adam. *The Problem of Trust*. Princeton: Princeton University Press, 1997

Slovic, Paul. "Perceived Risk, Trust, and Democracy." *Risk Analysis* 13.6 (1993): 675-681

Sproull L. & Kiesler S., 1991, *Connections*, The MIT Press, Cambridge, MA, 1991

Tygar & Whitten, 1996, "WWW Electronic Commerce and Java Trojan Horses", *Proceedings of the Second USENIX Workshop on Electronic Commerce*, 18-21 Oakland, CA 1996, 243-249

United States Council for International Business, 1993, *Statement of the United States Council for International Business on the Key Escrow Chip*, United States Council for International Business, NY, NY.

Wacker, J.,1995, "Drafting agreements for secure electronic commerce" *Proceedings of the World Wide Electronic Commerce: Law, Policy, Security & Controls Conference*, October 18-20, Washington, DC, pp. 6.

Walden, I., 1995, "Are privacy requirements inhibiting electronic commerce," *Proceedings of the World Wide Electronic Commerce: Law, Policy, Security & Controls Conference*, October 18-20, Washington, DC, pp. 10.