Security and Morality: A Tale of User Deceit

L Jean Camp
Indiana University
School of Informatics
+1-812-856-1865
ljcamp@indiana.edu

Cathleen McGrath
College of Business
Administration
Loyola Marymount University
+1-310-216-2045
cmcgrath@lmu.edu

Alla Genkina
UCLA
Information Studies
alla@ayre.org

ABSTRACT

There has been considerable debate about the apparent irrationality of end users in choosing with whom to share information, with much of the discourse crystallized in research on phishing. Designs for security technology in general, anti-spam technology, and anti-phishing technology has been targeted on specific problems with distinct methods of mitigation. In contrasts, studies of human risk behaviors argue that such specific targets for specific problems are unlikely to provide a significant increase in user trust of the internet, as humans lump and generalize.

We initially theorized that communications to users need to be less specific to technical failures and more deeply embedded in social or moral terms. Our experiments indicate that users respond more strongly to a privacy policy failure than an arguably more risky technical failure. From this and previous work we conclude that design for security and privacy needs to be more expansive in that there should be more bundling of signals and products, rather than more delineation of problems into those solvable by discrete tools. Usability must be more than the interface design, but rather integrate security and privacy into a trust interaction.

Categories and Subject Descriptors

Computers and Society

General Terms

Security, Management, and Experimentation

Keywords

Security, Trust, Trustworthiness

1. INTRODUCTION

1.1. Overview

In the first section of this paper we review the literature that inspired our trust experimentation. In the second section we describe our experiments. In the third section we discuss the results of the experimentation. In the fourth section we describe the potential implications of our results for the design of user interactions for risk communication.

Safe, reliable, and secure computing requires empowered users. Specifically users must be empowered to distinguish between trustworthy and untrustworthy machines on the network [13]. Of course, no machine that can be connected is perfectly secure. No home machine is without user information. To further complicate the transition, this evolution must occur in a dynamic widely-deployed network. The capacity of humans as security managers depends on the creation of technology that is designed with well founded understanding of the behavior of human users. Thus systems must not only be trustworthy but must also be identifiable as trustworthy. In order for this to happen we must root system development in an understanding of the cues that humans use to determine trustworthiness.

The efficacy of trust technologies is to some degree a function of the assumptions of human trust behaviors in the network. Note that the definition of trust in this project is taken from Coleman's [11] definition of rational actors' decision to place themselves in vulnerable positions relative to others in the hope of accomplishing something that is otherwise not possible. Its operational focus fits well with the computer science perspective. In contrast it is explicitly not the definition of trust as an internal state where confidence is expressed behavior as seen in [17].

Building upon insights that have emerged from studies on human-computer interaction and game theoretic studies of trust we have developed a set of hypotheses on human behavior with respect to computer-mediated trust. We then test these hypotheses using an experiment that is based on proven social science methods. We will then examine the implications for technical design of the confirmation or rejection of the hypotheses with the use of structured formal protocol analysis.

Technical security experts focus on the considerable technological challenges of securing networks, and devising security policies. These essential efforts would be more effective in practice if designs more systematically addressed the (sometimes irrational) people 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 people participate in and contribute to the security and trust of networks.

1.2 Theoretical Foundation

The study of network security is the study of who can be trusted for what action, and how to ensure a trustworthy network. This understanding must build upon not only the science and engineering of security, but also the complex human factors that affect when and how individuals are prepared to extend trust to the agents with whom they interact and transact - computers, people and institutions. This is a problem that has received much comment but little formal quantitative study [16, 25].

Humans appear to be ill suited as computing security managers. Arguments have been made for embedding security in the operating system from the psychological perspective [25]. In addition there is a continuous debate about making the network more trustworthy [10]. As technology becomes more complex, users develop simplified abstractions that allow them to make sense of complicated systems [36] but these flawed models may obfuscate vital security decisions. End-user security mechanisms may 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 argued in the case of applications [10].

Social science experiments provide insights for evaluating how trust mechanisms may succeed or fail when presented to the naïve user. That humans are a source of randomness is well-documented, and the problems of ‘social engineering’ well known. Yet the inclusion of the human behavior using tested axiomatic results is a significant extension to previous research on why security and trust systems fail [1].

The experiment described here was built upon the following theoretical construction of the problem. First, we narrow the larger question of security to the more constrained question of human trust behaviors. Second, we extract from the larger literature testable hypotheses with respect to trust behaviors. Third, we develop an experimental design where the trust behavior is a willingness to share information that give a basis for rejecting the testable hypotheses.

For this research, we use Coleman's [11] definition of trust that accounts for the rational action of individuals in social situations to structure the experimental situations which subjects will face. Coleman's definition of trust is operational and 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 off 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.

The view held by a number of researchers about trust is 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 such as confidence or reliance to denote the analogous attitude people may hold toward objects such as 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 versus confidence or reliance operate internally (See, for example, [28, 16]. Yet in terms of building mechanisms to create a trustworthy network we will investigate the way trust may be extended to both humans and objects. Note that there are disagreements with respect to the definition and examination of trust. Trust is a concept that crosses disciplines as well as domains, so the focus of the definition differs. There are two dominant definitions of trust: operational and internal.

Operational definitions of trust like the one we are using require a party to make a rational decision based on knowledge of possible rewards for trusting and not trusting. Trust enables higher gains while distrust avoids potential loss. Therefore risk aversion is a critical parameter in defining trust.

In the case of trust on the Internet operational trust must include both evaluation of the users intention – benevolent or malevolent, and the users' competence. Particularly in the case of intention, the information available in a physical interaction is absent. In addition, cultural clues are difficult to discern on the Internet as the face of most web pages are meant to be as generic as possible to avoid offense. One operational definition of trust is reliance [19]. In this case reliance is considered a result of belief in the integrity or authority of the party to be trusted. Reliance is based on the concept of mutual self-interest. Therefore the creation of trust requires structure to provide information about the trusted party to ensure that the self-interest of the trusted party is aligned with the interest of the trusting party. When reliance is refined, it requires that the trusted party be motivated to insure the security of the site and protect the privacy of the user. Under this conception trust is illustrated by a willingness to share personal information. Camp [8] offers another operational definition of trust in which users are concerned with risk rather than risk perception. From this perspective, trust exists when individuals take actions that make them vulnerable to others.

A second perspective on trust used by social psychologists, assumes that trust is an internal state. (e.g., [17]) From this perspective, trust is a state of belief in the motivations of others. Based on this argument, social psychologists measure trust using structured interviews and surveys. The results of the interviews can find a high correlations between trust and a willingness to cooperate. Yet trust is not *defined as* but rather *correlated with* an exhibited willingness to cooperate. This is in contrast to the working definition underlying not only this work, but also most of the research referenced herein. The definition of trust used here and the set of methods used to explore trust perfectly coincide and are based in the quantitative, game-theory tradition of experiments in trust in which trust is an enacted behavior rather than an internal state.

One underlying assumption is that, in addition to the technical, good network security should incorporate an increasingly systematic understanding of the ways people extend trust in a networked environment. Thus one goal of this experiment is to enable or simplify the design of systems enabling rational human trust behavior on-line by offering a more axiomatic understanding of human trust behavior and illustrating how the axioms can be applied. Therefore the goal of our experiment is to offer a way to embed social understanding of trust as exhibited in human action into the design of security systems. Yet before any concepts of trust are embedded into the technical infrastructure, any implicit hypotheses developed in studies of humans as trusting entities in relation to computers must be made explicit and tested. Then it is critical to illustrate by example how these hypotheses can be effectively applied to past technical designs.

This is a two-part research investigation. First, we test the hypotheses that are explicit in the game theory-based research on human trust behavior in the specific case of human/computer interaction. We test these hypotheses using standard experimental and quantitative methods, as described in the first methods section. Second, based on these findings, we examine the suitability of various distributed trust technologies in light of the findings of the first part of this study.

1.3. Hypothesis Development

We developed a core hypotheses under which the technologies of trust and the perspectives on trust from social science converge. Essentially in contrast to the assumption that individuals make increasingly complex decisions in the face of increasingly complex threats, social science suggests that people are simplifiers. The hypotheses at its core points to a common point of collision: technologists may embed in the design of trust mechanisms implicit assumptions that humans are attentive, discerning, and ever-rational. There are strong philosophical arguments that humans are simplifiers, and this implies that humans will use trust of machines to simplify an ever more complex world.

> **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 appearing to originate from malevolent human actions, on the one hand, and incompetence on the other.

According to this hypothesis 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 or human (or institutional) malevolence. Indeed, the determination to avoid risks without concern of their origination is a characteristic of risk technology.

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 any case. Yet a second examination yields a more complex problem space. This more complex design space in turn calls for a more nuanced solution to the problem of key revocation or patch distribution.

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 unpredictable, purely technical glitches;

cases where failures are attributed to technical shortcuts taken by human engineer; and thirdly cases where malevolence (or at least disinterest in another's situation) is the cause of harm. To briefly illustrate, a security breach that 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, the reaction might be more similar to a 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.

Our second hypothesis relates to the ability of individuals to make distinctions among different computers. Computers are of course, distinct, particularly once an operator has selected additional applications that will run on and policies that will govern the information on the site. Publications in social theory (e.g., [11, 31]) predict 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 [37].

Studies in human-computer interaction suggest that users, even those with considerable knowledge and experience, tend to generalize broadly from their experiences. Studies of off-line behaviors illustrate that such generalization is particularly prevalent in studies of trust within and between groups. Thus, positive experiences with a computer may generalize to the networked system (to computers) as a whole and presumably the same would be true of negative experiences. In other words, users may draw inductive inferences to the whole system, across computers, and not simply to the particular system with which they experienced the positive transaction. Do individuals learn to distinguish between threats or do they increase threat lumping behavior?

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

2. EXPERIMENTAL DESIGN

We collected data on computer users' responses to trustworthy and untrustworthy computer behavior by conducting real time experiments that measured individuals' initial willingness to conveying personal information in order to receive a service over the web, and then examined student responses to betrayals. A total of 63 students participated in the study. They were told that they were evaluating web pages as part of a business management class. . Students were shown one web site (elephantmine.net), then a second site (reminders.name).

The services offered over the Web sites appear to be life management services, that will require that individuals offer to provide information (e.g. birthday of your spouse, favored gifts, grocery brand preferences, credit card number). After participants viewed the web pages, they responded to a series of questions about their willingness to share information with the site. The survey determined the data the subjects were willing to provide to that domain. Our services portals are designed to be similar in interface but clearly different in source so that we can explore the question of user differentiation of threats.

This design has three fundamental components: trust, betrayal, trust. Subjects were told that they are evaluating e-commerce systems that will make their lives easier by managing gift-giving, subscription management, bill-paying, grocery shopping, and dry-cleaning etc. They were be asked their willingness to engage with such a company. Background information will included overall computer experience experiences. These questions included typical personal information as well as information about loved ones, daily habits, and preferences.

First we test the tendency for people trust to different machines as illustrated by a willingness to share information, as is consistent with referenced work. The two machines have different themes and different domain names. We showed that the machines are distinct types by clearly identifying the machine with visible labels (e.g. "Intel inside" and Tux the Linux penguin, vs. "Viao" and "powered by NT").

During the introduction of the second web page, there is one of two types of "betrayal". In the first, the betrayal is a change in policy that represents a violation of trust in terms of the intention of the agent. Here the students were shown a pop-up window announcing a change in privacy policy, and offered a redirection to a net privacy policy. In the second condition, "betrayal" represented a violation of trust in terms of a display of incompetence on the part of the agent. One segment of students were shown a betrayal that was another (imaginary) person's data being displayed on the screen. This illustrates a technical inability to secure information. After each "betrayal", we tested for more trust behaviors, again with trust behavior being defined as the willingness to share information.

3. RESULTS

The results of our experiment with users provides insight into our hypotheses regarding users' responses to violations of trust. Table 1 shows the results for the both conditions.

Table 1. Users' responses to betrayals

	Change in privacy policy (Malevolence)		**Display other users' private information (Incompetence)**	
Type of information	Proportion willing to share before	Proportion willing to share after	Proportion willing to share before	Proportion willing to share after
Your credit card number	0.16	.09 **	0.29	.13 **
Your Social Security number	0.03	0	0.03	0
Your year of birth	0.69	.59 ***	1	0.9
Your IM buddy list	0.22	.09 **	0.16	.13 ***
Your list of email contacts	0.13	.06 **	0.23	.13 ***
Your coworkers' names	0.44	.31 ***	0.42	0.52
Your friend's names	0.53	.34 ***	0.65	0.68
Your parents' names	0.47	.28 ***	0.58	.55 ***
Your family members' names	0.47	.28 ***	0.68	.61 ***
Your family members' birthdays	0.66	.47 ***	0.87	.68 **
Your family's wedding anniversaries	0.63	.47 ***	0.84	.68 ***
Your family members' shopping preferences	0.53	.38 ***	0.77	.71 ***

** p<.01
*** p<.001

In the first condition, there is a change in the privacy policy of the web page. We classify this as a violation of trust intention. According to the first 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.

In the second condition, participants saw that a fictional users' information was displayed when the webpage was opened. As shown in Table 1, after the technical error demonstrating incompetence, participants were less willing to share information, but by a smaller margin than in the first case of a change in privacy policy. Despite the fact that the technical failure indicated *an inability to keep information secure or secret or private*, the refusal to share future information far more dramatically decreased with the policy change.

The data above illustrates that we have explicitly rejected the hypotheses that all failures are the same, with respect to human-driven and technical failures.

The integration of the moral or ethical element is noticeably absent in security technology design even when there is an argument, without human interaction, that such a policy would be good security practice. For example, key revocation policies and software patches all have an assumption of uniform technical failure. 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 revocation lists are monolithic documents where the responsibility is upon the key recipient to check. Often, the key revocation lists only the date of revocation and the key. These experiments would argue that the cases of initial falsification, change in status, and lost device would be very different and would be treated differently. A search for possible fraudulent transactions or a criminal investigation would also view the three cases differently. Integrating the reason for key revocation may make human reaction to key revocation more effective and is valuable from a system as well as a human perspective.

The second hypothesis, that individuals develop mechanisms to evaluate web sites over time and enter each transaction with a new calculus of risk, cannot be supported by the evaluation. Each participant stated that they had at least seven years of experience of the web, including commerce. If the approach to a web site were one of careful updating of a slowly developed boolean function of risk, then the alteration in the second case arguably would have been less extreme. After all, the betrayal happens at the first site, not the second. So every participant should begin at the second site at exactly the same state as the first, assuming each differentiates web sites rather than reacting to experiences on “the net” as a whole.

Clearly there is no argument under which this data would support that argument. Individuals reacted strongly and immediately to the betrayal at the first site, despite being told that the first and second site were in no way related and were in fact competitors.

4. CONCLUSIONS

We have tested two hypotheses in human behavior that can serve as axioms in the examination of technical systems. Technical systems, as explained above, embody assumptions about human responses.

The experiments have illustrated that users consider failures in benevolence as more serious than failures in competence. This illustrates that distinguishing that security technologies that communicate state to the end user will be most effective if they communicate in terms that indicate harm, rather than more neutral informative terms. Systems designed to offer security and privacy, and thus indicating both benevolence and competence, are more likely to be accepted by users. Failures in such systems are less likely to be tolerated by users, and users are less likely to subvert such systems.

As the complexity and extent of the Internet expands users are increasingly expected to be active managers of their own information security. This has been primarily conceived in security design as enabling users to be rational about extensions of trust in the network. The truly rational choice is for security designers to embed sometimes irrational but consistent human behaviors into their own designs.

The consideration of people's responses to computers can be seen as drawing not only on the social sciences generally but specifically on design for values in its consideration of social determination. In the viewpoint of the social determinist, technology is framed by its users and adoption is part of the innovative process. That is to say, that designs are based on a post-hoc analysis of technologies after they have been adopted [16]. Beyond identifying flaws of security mechanisms we hope to offer guidance in the analysis of future systems. It would be unwise to wait until a security mechanism is widely adopted to consider only then how easily it may be undermined by "human engineering.”.

5. REFERENCES

[1] 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.

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

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

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

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

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

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

[8] Camp, L.J. *Trust & Risk in Internet Commerce*, MIT Press, 2000.

[9] Camp, L.J., Cathleen McGrath & Helen Nissenbaum, “Trust: A Collision of Paradigms,” Proceedings of Financial Cryptography, Lecture Notes in Computer Science, Springer-Verlag (Berlin) Fall 2001.

[10] 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.

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

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

[13] Computer Science and Telecommunications Board, 1999, *Trust in Cyberspace*, National Academy Press, Washington, D.C.

[14] 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

[15] 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.

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

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

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

[19] Golberg, Hill & Shostak, 2001 "Privacy, ethics, and trust" Boston University Law Review, V. 81 N. 2.

[20] 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.

[21] 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

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

[23] 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.

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

[25] Nikander, P. & Karvonen, "Users and Trust in Cyberspace. Lecture Notes in Computer Science, Springer-Verlag (Berlin) 2001.

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

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

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

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

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

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

[32] 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

[33] 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.

[34] 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.

[35] 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.

[36] Weick, K. "Technology as Equivoque: Sensemaking in new technologies" In Goodman, L. Sproull, eds. "Technology and Organizations. 1990.

[37] Weisband, S. & Kiesler, S. (1996). Self Disclosure on computer forms: Meta-analysis and implications. Proceedings of the CHI '96 Conference on Human-Computer Interaction, April 14-22, Vancouver.