

*Trust can be cultivated to  
enhance our personal and social lives and  
increase our social capital.*

# TRUST ONLINE

**T**rust matters. It allows us to reveal vulnerable parts of ourselves to others and to know others intimately in return. A climate of trust eases cooperation among people and fosters reciprocal care-taking. The resources—physical, emotional, economic—that would otherwise be consumed guarding against harm can be directed toward more constructive ends.

Here, we explore the nature of trust and how and where it flourishes online. We also seek to make sense of seemingly disparate perceptions. For example, some say the public is too trusting online; without thinking, people routinely download software likely to destroy important information or blithely engage in e-auctions or chat rooms with strangers. Others say the public does not trust enough, that people refrain, for example, from e-commerce under the mistaken belief that their financial transactions are not secure. How can we know if the trust we choose to give or withhold is warranted? Can we trust machines or other technological systems? How can those of us who create and maintain the technological infrastructure help establish a climate of trust?

Addressing such questions, we provide a conceptual framework for understanding trust, then offer 10 characteristics of online interaction that can help engineer trust online and distinguish

between trust in e-commerce activities and trust in online interpersonal interactions.

## **Conceptual Framework**

After killing Desdemona for her supposed betrayal, then realizing his grievous mistake, Shakespeare's Othello laments that he had loved not wisely, but too well. So it can be with trust. We can trust strangers we shouldn't and be betrayed by the people closest to us, including longtime friends, parents, children, and spouses. Conversely, we might misjudge and withhold our trust from those who wish us well. Both types of mistake—trusting too well and not well enough—can be costly.

But what is trust? The moral philosopher Annette Baier offered a useful starting point, writing: "One leaves others an opportunity to harm one when one trusts, and also shows one's confidence that they will not take it. Reasonable trust will require good grounds for such confidence in another's good will, or at least the absence of good grounds for expecting their ill will or indifference. Trust, then, on this first approximation, is accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one" [1].

Thus, we trust when we are vulnerable to harm from others yet believe these others would not harm us even though they could. In turn, trust depends on our ability to perform three types of assessments: the harm we might incur; the good will others have

---

toward us that might affect their efforts to protect us from harm; and whether or not harm that does occur lies outside the parameters of the trust relationship.

Common sense tells us that the barriers to trust are least inhibiting when the potential harm is minimal and the good will of the person(s) we trust is genuine (such as when loaning a small sum of money to a close friend). Conversely, barriers to trust occur when there is potentially significant harm and not much good will from the person(s) we trust (such as when loaning a large sum of money to a stranger). An example of whether or not such harm is outside the parameters of the trust relationship is when we trust engineers and builders to construct buildings that meet, say, current professional earthquake standards. If a major earthquake occurs—well beyond

the limits of what any building can withstand—and our building collapses, our trust in the engineers and builders is not betrayed, since protecting us from this harm was beyond their control. In other words, our trust in the designers of technology (or technological artifacts) is bounded by our understanding of the conditions under which the technology functions reliably and safely.

Betrayals of trust often end relationships. In Othello's case, even a suspected betrayal did so. Sometimes a lesser violation is better viewed as a breach than as an outright betrayal. Breaches may arise from small harms or small mistakes in judgment of another's good will; they can also be repaired or occur outside the core parameters of a relationship. For example, not long ago, customers who regularly read Amazon.com's online book recommendations assumed they were

editorial content written by Amazon.com staff. A breach of trust resulted when it was revealed that publishers sometimes purchased spots for their books in this recommendation system [11]. But this breach did not occur at the core of the commercial relationship Amazon.com has with its potential customers and Web site users. Thus, although we might guess the breach did not negatively affect the commercial side of Amazon.com's site, it left lingering doubts about the validity of practically any of its many recommendation systems. Over time, if Amazon.com refrains from this business practice, and if no further breaches are revealed, the online book-buying public will likely recover its trust in the company's editorials.

In contrast, consider the recent history of TRUSTe, a nonprofit organization whose mission is to build user trust in the Internet by promoting the principle of disclosure. Following an online organization's request, TRUSTe audits the organization's Web site. If it meets certain minimal criteria, TRUSTe

able to experience good will, extend good will toward others, feel vulnerable, and experience betrayal. These psychological states, in turn, depend on consciousness and agency. Without veering too far into philosophical argument, it is clear that human beings have consciousness and agency. The same cannot be said of a technological system in and of itself. We may speak of autonomous agents, goal-directed algorithms, intelligent machines, and the like in reference to behavior. But technological artifacts have not yet been produced in substance and structure that warrant in any stringent sense the attribution of consciousness or agency. People trust people, not technology.

We recognize that at least two critiques can be levied against this position: The first is that in our common language, people often use the term "trust" broadly, referring to expectations for technologies (let alone physical phenomena). In this sense, we say we trust that brakes will stop a car (or the sun will rise again). Indeed, the Computer Science and



## PEOPLE TRUST PEOPLE, NOT TECHNOLOGY.

allows the site to display the TRUSTe seal of approval and maintain a TRUSTe certification. For example, about a year ago, TRUSTe conducted an initial inquiry into the behavior of RealNetworks' RealJukebox, which had been distributing software that surreptitiously gathered personal data from users' hard disks. TRUSTe responded by claiming such actions were beyond the scope of the TRUSTe audit, because the RealJukebox software worked only indirectly through a Web site visit. In turn, many people believed TRUSTe had evaded its responsibility, pointing out that such a breach of trust occurred on the heels of a handful of other breaches [9]. If it's true that such breaches of trust occur repeatedly at the heart of TRUSTe's mission, it will be difficult if not impossible for TRUSTe to recover the trust of the online public.

*What or whom can we trust?* Online interactions represent a complex blend of human actors and technological systems. In light of this complexity, with what or whom can we meaningfully speak of building trust relationships? The system? Its developers? Web site designers? Online organizations? Other users?

To answer, recall that trust exists between entities

Telecommunications Board, in its thoughtful 1999 publication *Trust in Cyberspace* adopted the terms "trust" and "trustworthy" to describe systems that perform as expected along the dimensions of correctness, security, reliability, safety, and survivability [10]. However, in our view, this broad use of the term trust introduces unnecessary confusion [6]. Imagine, for example, that some technical aspect of the Internet (such as a remote server) fails to perform as expected, resulting in some harm, perhaps the loss of time, information, or privacy. Here, we have a simple case of technical failure. Moreover, by trying to invoke the idea of a trust violation in this context, we could mistakenly seek moral remedies, to, say, understand how an assessment of good will is faulty or how to cultivate increased good will, when moral issues were not at stake to begin with. In short, equating a technical failure with a violation of trust conflates both the non-moral and moral sources of the problem and diverts important resources from identifying meaningful remedies.

The second critique grants that it is reasonable to speak of relying on (but not trusting) simple machines or even a single computer in this way. But it is then argued that diverse, complex, and self-

evolving systems like the Internet actually create an “atmosphere” of trust (or lack thereof). Therefore, these complex systems should be viewed as valid participants in trust relationships. As the argument continues, an end user must first trust in that atmosphere—the technology and the human community combined—and only then is positioned to trust in any particular online interaction with other people. We agree that the environment (physical or online) in which people interact has decisive effects on a person’s desire and ability to participate in trust relationships. But such effects arise not because the environment enters into the trust relationship but because people frequently draw on cues from the environment to ascertain the nature of their own vulnerabilities and the good will of others.

While people have sometimes applied the term trust too broadly in the technological realm, they have applied it too broadly in the social realm as well. For example, at a recent human-computer interaction conference, a colleague attributed failure of collaboration between two remote work groups to the “problem of establishing trust” among the participants in the two groups. But further dialogue revealed that a major problem arose because members in one group couldn’t identify the official authority in the other group and were frustrated seeking the appropriate person to sign off on their work. Is this failure to communicate a problem of trust? We think not. Thus, as trust in online interactions becomes increasingly central to our public discourse (and covered in the popular media), it is increasingly important that we not conflate trust with other important aspects of social interaction.

### **Engineering Conditions for Cultivating Trust Online**

That we trust people, not technology, is not to say that technology is value-neutral. Rather, technologies in general, and computer technologies in particular, provide “suitabilities” that follow from features of the technology. That is, a given technology is more suitable for certain activities and more readily supports certain values while rendering other activities and values more difficult to realize [2]. For example, a hammer is suited for driving nails but makes a poor ladle or pillow. An online calendar system that displays individuals’ scheduled events in detail readily supports accountability within an organization but makes privacy difficult.

How can we engineer technology that cultivates the conditions for trust online? In order to answer, we outline the following 10 trust-related characteristics of

online interaction we find helpful in our analysis and design work.

*Reliability and security of the technology.* Given the best professional practice of the day, there is much about this technology that is not yet reliable or secure. For example, inspection alone cannot determine whether code is safe. Nor is it possible to know for certain that some third party is not impersonating a Web site. Thus, end users must decide whether or not to trust an online environment in which certain vulnerabilities are unknown, even to the most knowledgeable individuals.

*Knowing what people online tend to do.* Risks abound online—or so we are told. Users fear viruses, hackers, and other users in disguise. But how prevalent are these risks? What percentage of users engage in hacking? How often does it harm the typical user? How much harm does it do? What percentage of users masquerade online? Do people masquerade everywhere online or only in limited venues like online chats? Is the deception online any greater than its counterpart in real life? The point is that we have only limited accurate information about how great are the risks online and how frequently they occur. Thus, the problem for establishing trust online is how to do so in light of enormous uncertainty about both the magnitude and the frequency of potential harm.

*Misleading language and images.* Think about the word “secure.” If someone tells us that a serial killer is locked up in a secure prison, we would rightly expect the chance of an escape to be extremely unlikely. But when designers tell us there is a “secure connection” for the http protocol, we do not know what level of security we can expect. Note too that the padlock-and-key icon in Netscape Communicator conveys a type and level of security that is not comparable to our experience with the icon’s physical counterparts. Consider again the TRUSTe seal of approval, which would seem to offer an independent assessment of how well an online organization ensures privacy. On closer inspection, TRUSTe grants its seal to any organization that follows its own posted privacy policy, which sometimes does not ensure privacy at all. The situation would be akin to a hotel garnering a five-star rating simply by promising not to guarantee its customers good service and then faithfully keeping its promise. In each case, the larger issue recalls our earlier proposition: Trust depends not only on assessing harm and good will but what to reasonably expect of the technology. Yet designers routinely use misleading language and images to convey to users greater reliability and security in the technology than is warranted.

*Disagreement about what counts as harm.* We noted the difficulties of assessing the harms that may occur from breaches of trust. This issue concerns the accuracy of information about events. However, even when there are shared understandings about what really happened in a particular event, there may still be legitimate disagreements about whether and to what extent harm has occurred. For example, in research on adolescent conceptions of property and privacy, some adolescents considered accessing another's computer file without reading the contents a privacy violation; others did not [3]. Adolescents who did not view the event as a privacy violation reasoned in one of two ways: Either that no harm had occurred (for example, [Accessing a computer file without reading it] "is just an act of defiance ... All you're doing is faking them out, and you're not hurting them"); or that no rights had been violated (for example, "I don't think you're invading their privacy because you haven't actually read it [the computer file], you've just proven to yourself that you could read it if you wanted to"). Without debating the normative position here, what counts as harm in online interactions may not have broad societal agreement.

*Informed consent.* Should organizations be allowed to put cookies on users' machines? Should they be allowed to track the mouse movements of individual users visiting Web sites? Should they be allowed to generate personal buying patterns of their online customers, then offer them individualized promotions? Should they be allowed to share customer profiles with one another? How about financial and medical profiles? These and other questions are best answered by individual users, not by CEOs, marketing executives, or system designers. In turn, informed consent provides a means for garnering each user's answer.

Informed consent involves telling users of the potential harm or benefit of an online interaction and giving them the explicit opportunity to consent or decline to participate in the interaction. To date, informed consent is woefully understudied by the online community and underused as a means of cultivating trust online. In an effort to provide some formal guidance, there has been some movement toward developing criteria and design principles for implementing informed consent online [5].

*Anonymity.* Anonymity refers to the absence of identifying information associated with an interaction [8]. Compared to physical interactions, online interactions allow for both greater and lesser amounts of anonymity. For example, in a physical cash transaction, no paper or digital trail connects

the consumer to the purchase or links purchases over time. But there is face-to-face contact that, in, say, a small town, can noticeably decrease one's privacy. In online interactions, the lack of face-to-face contact leads to greater anonymity, at least in interpersonal interactions. Yet in e-commerce, each consumer leaves behind a long digital trail, including name, credit card numbers, mailing address, and buying patterns. Without trying to explicate the complexity of how anonymity manifests online, we can say that on the one hand, anonymity can erode a climate of trust by making assessments of potential harm and good will of others more difficult. On the other, if we focus on protecting ourselves from the potential harm and ill will of others, then anonymity can help cultivate a climate of trust by putting in place greater safeguards.

*Accountability.* High degrees of anonymity provide significant challenges for accountability [7]. After all, if you do not know the person with whom you are interacting and are unable to track the person to a location, there are fewer incentives for a stranger to behave with good will. But as we increase accountability (and seek to minimize potential harm), we often decrease anonymity (and increase violations of privacy and undermine personal autonomy). Careful attention needs to focus on the balance between anonymity and accountability, so in our efforts to engineer trust online we do not unduly override other important human values.

*Saliency of cues in the online environment.* The presence or absence of cues embedded in the online environment can alter the conditions needed for trust. For example, if in an AIDS bulletin board all the posted information was stripped of its professional sources, users would be unable to distinguish whether the source of some "cutting-edge" treatment is a layperson or a medical doctor. Thus, status cues can increase user confidence in the source and quality of information. Of course, the absence of status cues opens channels of communication in otherwise hierarchically oriented relationships. Here again, designers must balance trust with other human values.


*Insurance.* Insurance refers to social arrangements in which there is a "promise" to compensate individuals for future harm if it occurs. In e-commerce, insurance is often offered in terms of financial compensation (such as by fully covering the cost of a credit card purchase that goes awry) or some other arrangement (such as seeking to recover data destroyed by mistake). In evaluating insurance, users should consider not only the adequacy of the compensation but the ease of obtaining it as well.

*Performance history and reputation.* In order to judge one's vulnerability online, assess performance history, including direct past experiences with the party in question, along with the reported experiences of others. Indeed, organizations sometimes offer their online customers an accounting of such experiences. For example, to help establish a climate of trust, one well-known Web site has provided quantitative statements of its performance history to new customers, including: "None of our three million customers has reported fraudulent use of a credit card as a result of purchases made at ..." Performance histories, in turn, create reputations, through, say, qualitative accounts of performance, such as "My friend says this online company is great to deal with."

Some of these characteristics are transitory, emerging because users are relatively inexperienced with the technology. For example, the absence of perfor-

trust online. In doing so, it is important for designers to distinguish two overarching contexts for trust online: e-commerce and interpersonal relationships.

*E-commerce.* In online commercial transactions, we are vulnerable to trust violations in two ways: loss of money and loss of privacy. Certain characteristics of online technology, such as those involving security, anonymity, accountability, and performance history, can make it difficult for users to determine the potential for both financial harm and the good will of the organization they're dealing with. Note that to buy something, consumers often rely on search engines to point them to a particular organization. Often, they have never even heard of the organization before, live hundreds or thousands of miles from its location (that, for all they know, could be in a one-room storage unit), and are serviced, if at all, by salespeople with whom they



THE SITUATION WOULD BE AKIN TO  
A HOTEL GARNERING A FIVE-STAR RATING  
SIMPLY BY PROMISING NOT TO GUARANTEE  
ITS CUSTOMERS GOOD SERVICE AND THEN  
FAITHFULLY KEEPING ITS PROMISE.

mance history is more an artifact of the technology's novelty than it is a deep technological feature. Over time, as performance histories develop, users are better positioned to assess the magnitude and likelihood of potential harm. Other technology features affecting trust relationships are more deeply tied to a particular technology's structure or stem from persistent social conditions surrounding the technology's use. Users can expect these features to continue well beyond the initial period of integration. For example, anonymity (and the absence of face-to-face interactions online) will continue to challenge their assessment of the potential good will of strangers.

### **Online Context**

Whether the characteristics are transitory or structural, the goal should be to engineer technology that more suitably cultivates the conditions for

never actually speak. Without something more, few of us would trust such a financial transaction. Insurance is that something more and is as simple as it is pervasive (now used by credit card companies and a number of online organizations seeking to limit the financial risk to their customers of fraudulent commercial transactions). Interestingly, the more confidence users have in a well-designed mechanism limiting their financial risk, the less trust they must demand of the commercial party in question.

As for privacy, technology today allows organizations to collect personal customer and client information and share it with one another. The resulting privacy violations trouble many and are leading to the adoption of governmental regulations. Technology should be designed to minimize such violations. One example is the eGenie Web site, which recommends movies, books, music, events, and TV shows based on

user profiles. and claims to allow users to remove their data from its recommendation systems should they decide they no longer want to participate. Designing systems that allow for informed consent is crucial.

**Online interpersonal interactions.** Violations of trust in online interpersonal interactions also make us vulnerable psychologically, producing, say, hurt feelings or embarrassment. As in other interpersonal relationships, there are no guarantees. We cannot, for example, take out an insurance policy to protect ourselves from psychological harm should we experience betrayal in a friendship.

Anonymity is double-edged in interpersonal relationships. On the negative side, online anonymity can limit the depth of interpersonal interactions insofar as we engage in a singular means of expression (written). On the positive side, online anonymity represents an opportunity for important interpersonal interactions. For example, a gay teenager in an intolerant family or community might rely on the anonymous characteristics of the Web to find like-minded peers online. Thus, in the interests of enhancing interpersonal trust, we need to develop tools that allow users to control what personal information is made known to others online.

As an example, when the city government in Santa Monica, CA, wanted to resolve friction between wealthy beachfront property owners and the homeless living on the beaches, it acknowledged the difficulty these two groups would have engaging in face-to-face discussion. The city utilized a community network to encourage online dialogue between the groups [12]. One way to appreciate the success of this forum is that it helped foster communication and good will, and some measure of trust, by diminishing the saliency of the barriers of social class (such as that many of the wealthy rejected face-to-face interactions with the homeless due to the latter's lack of personal hygiene).

## Conclusion

How can we design and use these technologies wisely and ethically to enhance our personal and social lives? We have broached this question through a discussion of the enduring human value of trust. Perhaps the greatest difference between trust online and in all other contexts is that when online, we have more difficulty (sometimes to the point of futility) of reasonably assessing the potential harm and good will of others, as well as what counts as reasonable machine performance. That is why people can engage in virtually identical online interactions, yet reach widely disparate judgments about whether the interactions are trustworthy.

More broadly, our approach—from a conceptualization of the enduring human value of trust to its working out in the technological arena—fits within the emerging multidisciplinary field called “value-sensitive design.” The result is technology that accounts for human values in a principled and comprehensive manner throughout the design process [2, 4]. As more work is conducted under this rubric, we can look forward to the field taking shape in terms of scope, methods, criteria, and metrics. ■

## REFERENCES

1. Baier, A. Trust and antitrust. *Ethics* (Jan. 1986), 231–260.
2. Friedman, B., Ed. *Human Values and the Design of Computer Technology*. Cambridge University Press, New York, 1997.
3. Friedman, B. Social judgments and technological innovation: Adolescents' understanding of property, privacy, and electronic information. *Comput. Hum. Behav.* 13, 3 (1997), 327–351.
4. Friedman, B. *Value-Sensitive Design: A Research Agenda for Information Technology* (contract no: SBR-9729633). National Science Foundation, Arlington, VA, 1999; see [www.ischool.washington.edu/vsd/](http://www.ischool.washington.edu/vsd/).
5. Friedman, B. Felten, E., and Millett, L. *Informed Consent Online*. The University of Washington, 2000.
6. Kahn, P., Jr. and Turiel, E. Children's conceptions of trust in the context of social expectations. *Merrill-Palmer Quart.* 34 (1988), 403–419.
7. Nissenbaum, H. Accountability in a computerized society. *Sci. Engin. Ethics* 2 (1996), 25–42.
8. Nissenbaum, H. The meaning of anonymity in an information age. *Inform. Soc.* 15 (1999), 141–144.
9. RealNetworks is target of suit in California over privacy issue. *The New York Times* (Nov. 9, 1999), C16.
10. Schneider, F., Ed. *Trust in Cyberspace*. National Academy Press, Washington, DC, 1999.
11. Rosman, K. Booking plugs on Amazon.com. *Brill's Cont.* (Apr. 1999).
12. Van Tassel. Yakety-Yak, do talk back!: PEN, the nation's first publicly funded electronic network, makes a difference in Santa Monica. In *Computerization and Controversy: Value Conflicts and Social Choices, 2nd Ed.*, R. Kling, Ed. Academic Press, Boston, MA, 1991, 547–551.

---

**BATYA FRIEDMAN** (batya@u.washington.edu) is an associate professor in The Information School and an adjunct associate professor in the Department of Computer Science and Engineering at the University of Washington in Seattle.

**PETER H. KAHN, JR.** (pkahn@u.washington.edu) is a research associate professor in the Department of Psychology at the University of Washington in Seattle.

**DANIEL C. HOWE** (dchowe@u.washington.edu) is a Ph.D. student in The Information School at the University of Washington in Seattle.

---

This work is supported in part by a National Science Foundation Grant (SES-0096131).

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.