Toward an Ethics of Persuasive Technology

Ask yourself whether your technology persuades users to do something you wouldn't want to be persuaded to do yourself.

DANIEL BERDICHEVSKY AND ERIK NEUENSCHWANDER

Technologies have always influenced our lives and how we lead them, but for the most part, their effects on our attitudes and behaviors have been incidental, even accidental. For example, automobiles and highways helped create the American suburbs, but they were not invented with the intent of persuading tens of millions of people to commute to work every day. Early computer spreadsheets gave us the number-crunching abilities needed to model future financial decisions, but did not advise us to take particular actions or reward us for what their designers might have viewed as "good" choices.

Likewise, there have always been human persuaders in society, masters of rhetoric capable of changing our minds, or at least our behaviors. Obvious examples of persuaders abound—cult leaders, mothers, car salesmen. Teachers, too, are persuaders of an invisible yet fundamental sort, altering the attitudes of their students day by day.

Persuaders often turn to technology to amplify their persuasive ends, as when Adolf Hitler literally amplified his voice using a megaphone to sway the German masses toward war, genocide, and a new social order. Though the megaphone facilitated Hitler's persuasion, on its own, it could not have persuaded anyone to do anything.

Likewise, a television can display commercials or influential after-school specials, but only if someone is transmitting them. Stripped of a signal, the television shows only static.

Only recently have technologies emerged that are actively persuasive in their own right, artifacts created primarily to change the attitudes and behaviors of their users. The study of such technologies is called "captology" (see the Introduction to this special section).

What if home financial planning software persuaded its users to invest in the stock market? And what if the market then crashed, leaving the users in financial ruin? Or, more subtly, what if the makers of the software arranged with certain companies to "push" their particular stocks? Would such designs differ in a morally relevant

The Principles of Persuasive Technology Design

I

The intended outcome of any persuasive technology should never be one that would be deemed unethical if the persuasion were undertaken without the technology or if the outcome occurred independently of persuasion.

II

The motivations behind the creation of a persuasive technology should never be such that they would be deemed unethical if they led to more traditional persuasion.

III

The creators of a persuasive technology must consider, contend with, and assume responsibility for all reasonably predictable outcomes of its use.

IV

The creators of a persuasive technology must ensure that it regards the privacy of users with at least as much respect as they regard their own privacy.

V

Persuasive technologies relaying personal information about a user to a third party must be closely scrutinized for privacy concerns.

VI

The creators of a persuasive technology should disclose their motivations, methods, and intended outcomes, except when such disclosure would significantly undermine an otherwise ethical goal.

VII

Persuasive technologies must not misinform in order to achieve their persuasive end.

VIII

The Golden Rule of Persuasion

The creators of a persuasive technology should never seek to persuade a person or persons of something they themselves would not consent to be persuaded to do.

Figure 1. Ethical principles of persuasive design

way from stockbrokers who encourage their clients to buy the stocks that earn them bonus commissions [6]? They do, though in unexpected ways. That's why our exploration of the ethics of persuasive technologies seeks to begin establishing a first set of principled guidelines for their design and implementation (see Figure 1). Doing so requires us to apply to this new domain a number of questions (and answers) associated with earlier work in the ethics of persuasion and in the ethics of technology—especially computers. Until now, no one has looked specifically at the convergence of these fields.

Articles about ethics are often peppered with jargon; in this way, ethics is a lot like computer science. But we avoid specialized terms except where they meaningfully enrich our discussion—and even then, we define them in context.

We view persuasion as an intentional effort to change attitudes or behavior and technology as the directed application of abstract ideas. Passive technological media, such as megaphones and billboards, facilitate persuasion without altering their pattern of interaction in response to the characteristics or actions of the persuaded party. Active persuasive technologies, however, are to some degree under the control of or at least responsive to the persuaded party. Or should be. The appearance of control may suffice for creating a persuasive experience, but if this appearance is not backed up by reality, the designer runs afoul of our accuracy principle.

Between active persuasive technologies and passive technological media are what we term "structural persuasive technologies," such as carpool lanes. While they are interesting examples of passive technologies that are not media, our focus here is on active persuasive technologies.

We refer to ethics as a rational, consistent system for determining right and wrong, usually in the context of specific actions or policies. The creation of a persuasive technology is such an action. Admittedly, there are almost as many possible systems of ethics as there are ethicists.

In strict deontological ethics, certain standards of conduct can never be broken, even when obeying them might cause someone grief or when breaking them once or twice (as in telling a white lie) might bring a person happiness. Taking a very different approach, act-based utilitarians evaluate the ethics of any action by gauging its consequences with respect to a particular criterion—usually human happiness or well-being. This approach can be thought of as "pro and con" ethics. A comfortable middle ground is rule-based utilitarianism, in which we stipulate ethical rules only if *always* following them results in more compelling benefits.

At first glance, many of the design principles we postulate for persuasive technology might seem deontological, but for the most part, they stem from a rule-based approach. You might also regard them as risk factors. The more of them a designer violates, the greater the risk the resulting design will be ethically problematic.

Throughout the article, we refer to the ethics of persuasive technology, instead of to the ethics of captology. Captology is the study of persuasive technology (see Figure 2), just as zoology is the study of animal species and political science the study of government. Zoology and political science are themselves neither ethical nor unethical, though the treatment of zoo animals or the behavior of government officials might be valid areas for ethical inquiry.

To explore ethical issues in persuasive technology in a compelling way, we annually solicit from students "dark side" designs, that is, applications of persuasive technology with troubling ethical implications. (Two of these fictional but provocative designs are described in the sidebar From the Dark Side.)

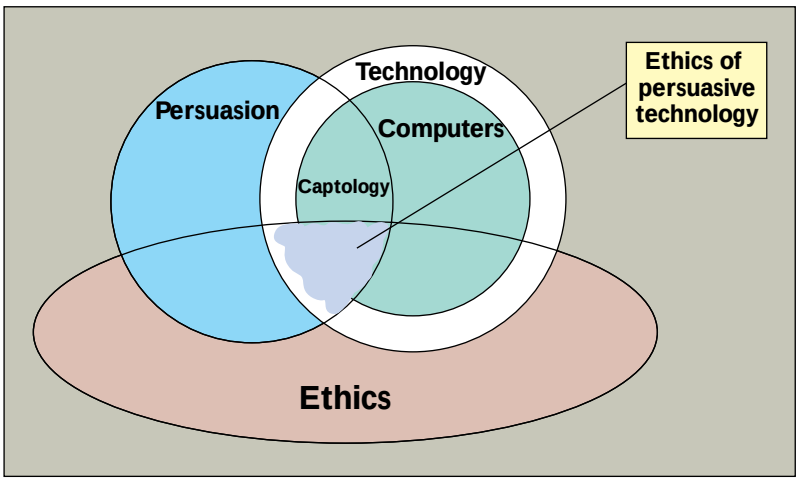

Figure 2. Convergence of ethics, persuasion, and technology. Ethical concerns extend beyond persuasive computers to all forms of persuasive technology—from the simply structural to the complex and cybernetic.

Uneasy Ethical Ground

Persuaders have always stood on uneasy ethical ground. If a serpent persuades you to eat a fruit, and if in eating it you cause humanity moderate distress, does culpability fall upon you or upon the serpent? Ethicists have struggled with such questions for thousands of years—and so has every persuader with a conscience.

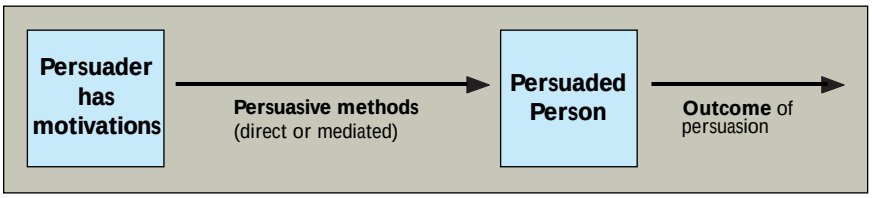

Figure 3. Framework for evaluating the ethics of a persuasive interaction in a traditional persuasive context.

Persuasion apparently distributes responsibility between the persuader and the persuaded. In most simple cases, where one person is persuading another, we agree with the ethicist Kenneth E. Andersen who has argued that all involved parties share full moral accountability for the outcome [1]. In our view, if Brian convinces Fannie to murder Jeff, Brian and Fannie are each responsible for the murder. Fannie is no less responsible because someone talked her into it; in the end, she made the choice herself. Some parties might dismiss Brian as

From the Dark Side

Design name: The Missionary

Purpose: Facilitate conversion of people in a particular region to a new religion

How it works: Proselytizers distribute glowing necklaces or other trinkets to newly converted believers. These trinkets can be tracked by a central computer and must be "recharged" regularly at a place of worship to maintain their glow. If not recharged in a reasonable time, their signal to the central computer begins to fade, and missionaries are quickly sent out to restore their owners' faith.

Design name: My Secret Pal

Purpose: Persuade children to divulge their secrets, so parents can take better care of them

How it works: A doll or toy is designed to be able to tell children its secrets. Children naturally reciprocate. Later, their parents can search a record of their children's secrets, using the information to adjust their parenting strategies accordingly.

only an accessory to the murder, but none would dispute that he bears some degree—perhaps a large degree—of responsibility for it.

Analyzing the ethics of any specific persuasive act requires a systematic approach, beginning with a breakdown of standard persuasion and eventually encompassing persuasive technologies as needed. To support this approach, we propose a framework for analyzing acts of persuasion according to their motivations, methods, and outcomes—intended and unintended. Our development of the framework begins with the basic relationship of a persuader and a person being persuaded (see Figure 3). In these instances, while a persuader may still use technologies like megaphones and billboards to convey the persuasive message, we ultimately l ook only at the two parties when distributing responsibility.

However, our focus is on technologies created with the intention to persuade—sometimes called "endogenously" persuasive technologies. They differ from technological media in that they are actively persuasive intermediaries between the persuader and the persuaded person. Unlike billboards, they interact dynamically with the objects of their persuasion (see Figure 4).

The framework of motivations, methods, and outcomes can be applied in evaluating the ethics of a persuasive act in either case, but the introduction of an actively persuasive technology requires the separate attribution of motivations to the designer and of the persuasive intent to the technology. Oddly, but meaningfully, the technology is both a method and the direct executor of persuasive methods.

We must also consider whether technology alters or even shares in the distribution of responsibility for the intent, methods, and end result of a persuasive act. To explore this possibility, we turn to an example from the study of computer ethics. In 1991, an automated Volkswagen factory automatically unexpectedly sped up its production facilities to move many more cars than the system could handle, sending many of them rolling off the assembly line straight into a wall [2].

Who is liable for the smashed cars and for the collateral damage to the factory? We can begin to resolve such questions once we accept as a reasonable assumption that human beings are free moral agents, though influenced by biology.[1] While we sometimes act predictably, predictability does not render us automata. We have intentionality, or at least a compelling enough illusion of it, that for all intents and purposes we ought to accept it as real. To date, computers have demonstrated neither the capacity to form their own intentions nor the ability to make their own choices. By any sensible standard, therefore, they are not free moral agents [4]—so when computers make serious mistakes, their programmers are often the first people blamed, users second, and Mother Nature third. The computer itself gets off easy.[2]

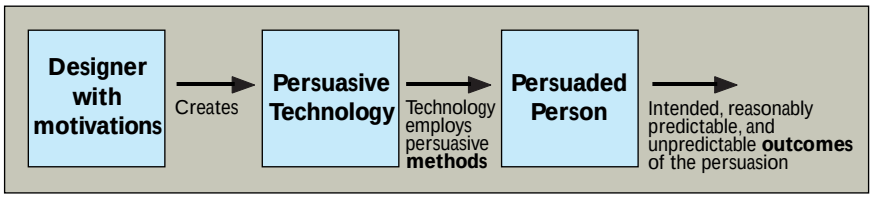

Figure 4. Framework for evaluating the ethics of the more complex interaction of persuader, persuasive technology, and the party or parties being persuaded.

Similarly, we cannot realistically attribute responsibility for the persuasive act to the persuasive technology. If a slot machine with a compelling multimedia narrative entices people to gamble away their savings, the slot machine itself is not at fault. Nor does the slot machine deserve credit for making the experience of gambling more entertaining. Rather, responsibility for the computerized machine's built-in motivations, methods, and outcomes falls squarely on its creators and purchasers; responsibility for the gambler's choice to gamble is distributed to both these parties—and to the gamblers themselves—just as if a human being were doing the persuading.

The major difference between persuasion through active technology and through traditional person-to-person relationships and interactions is not motivation, since the persuader still intends to persuade, presumably for the same reason or outcome, and since the persuaded person still undertakes or experiences that outcome. Our ethical scrutiny of persuasive technology has to center on the methods employed in the persuasion itself.

[1]Certain biochemical agents, such as antidepressants, can cause drastic changes in a person's state of mind and chosen behaviors. This example shows that people's intentions emerge to some extent from their biological makeup. But this extent need not be deterministic, nor is it predictable. Because people feel to themselves like free beings, as ethicists and as everyday people, they ought to function within this framework of perceived freedom.

[2]People sometimes treat their computers as if they were free moral agents, getting angry at a program when it crashes or at a printer when it jams. But is a computer displaying child pornography "dirty" or morally reprehensible? No, because the computer has no sense of what is appropriate beyond the instructions given it by human-driven programs.

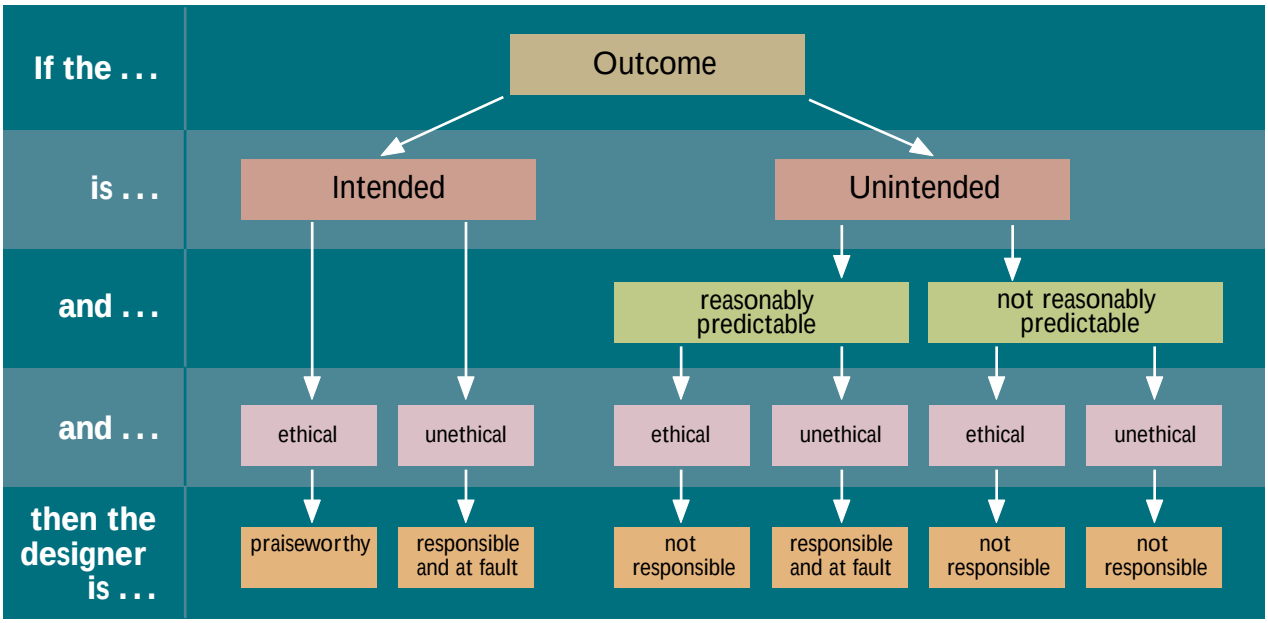

Figure 5. Flow chart clarifying the levels of ethical responsibility associated with predictable and unpredictable intended and unintended consequences.

Motivations vs. Intent

The motivations underlying a persuasive act and the intent of that persuasive act are not the same. To figure out the motivation of a persuader, ask yourself, Why is this person persuading me or someone else to do something? Consider three people sharing a common intent to persuade a stranger to eat more fruits and vegetables (see King et al.'s "The Landscape of Persuasive Technologies" in this issue). One might be motivated by a desire to increase the stranger's quality of life, the second by a mandate to increase revenue for the family farm, and the third by a secret hope the stranger will eat a bad fruit and become sick to the stomach. The persuasive intent is constant, even as the motivation varies in an ethically relevant way. The first is clearly laudable, the third problematic. The second falls in a more neutral zone, in which cultural context (discussed later) regarding commercialism and the other factors in our framework of persuasive methods and outcomes grow in relative importance.

We must also look at the methods through which a persuader persuades. If a person were to convince a stranger to eat fruit by playing on exaggerated fears, we might judge these methods unethical, even if the motivations appeared to be laudable.

The methods employed by persuasive technology are similar to those employed by persuasive people. For example, humans can persuade through flattery. Recent research has shown that computers can flatter too [3]. Humans can also persuade through conditioning, by rewarding and punishing desirable and undesirable behaviors. So can computers. However, technologies embed these methods in a new and compelling context. For instance, while humans can persuade through basic role playing, computers permit simulations to achieve unprecedented complexity, realism, and persuasive potential (see Khaslavsky et al.'s "Understanding the Seductive Experience" in this issue). Such differences are why we need to reconsider the implications for the ethics of traditional persuasive methods when these methods are undertaken by technologies instead of by humans.

We must also evaluate the ultimate outcome of the persuasive act—the ethics of what the persuaded person is persuaded to do or think. If something is unethical for you to do of your own volition, it is equally unethical to do when someone persuades you to do it. What about unintended outcomes? Suppose the stranger proved severely allergic to and died after ingesting a kumquat. Few people are allergic to kumquats, so this unfortunate but unintended outcome would not be considered reasonably predictable, nor would the persuader be held responsible for the outcome. However, if this were a common allergy and the ensuing reaction thus reasonably predictable, the persuader would have to be called to account. Later, when we discuss simulation as a persuasive method, we assert that designers of persuasive technologies should be held responsible only for reasonably predictable outcomes (see Figure 5).

How sensitive should designers and programmers be to the ethics of the persuasive technology they design? Imagine that someone persuading through a

more traditional medium, like a billboard, hires an artist to paint it. The artist seems to be in a position analogous to that of a programmer contracted to devise a persuasive program. However, because programmers create active persuasive agents, they are even more accountable than the artist for the persuasive nature of their work, and especially for the persuasive methods these agents employ. Programmers should never be reduced to mercenaries doing any kind of work for hire without sharing in the responsibility for it—a principle that comes across clearly in the ACM Code of Ethics (see www.acm.org./constitution/code.html) and that cannot be overstated.

Given this framework of motivations, methods, and outcomes, we can establish the first three of our principles for future persuasive-software design:

- The intended outcome of any persuasive technology should never be one that would be deemed unethical if the persuasion were undertaken without the technology or if the outcome occurred independent of persuasion.
- The motivations behind the creation of a persuasive technology should never be such that they would be deemed unethical if they led to more traditional persuasion.
- The creators of a persuasive technology must consider, contend with, and assume responsibility for all reasonably predictable outcomes of its use.

Certain acts of persuasion may not be practical without technology. For instance, it would be difficult to persuade someone through conventional means to maintain the proper pulse rate during exercise, but a simple biofeedback monitor can intervene appropriately. Implementation of persuasive technologies in such domains, where conventional persuasive techniques are difficult at best, calls for heightened ethical scrutiny. Even so, it is still possible to ask in a thought experiment whether it would be ethical for a person to attempt the persuasion without technology.

The Dual Privacy Principles

Human persuaders often exploit information about the people they persuade, and is an important reason why friends and family members can be the best people to persuade any of us to do something. They know more about us and our needs, including information we might prefer to keep private, and can adapt their persuasive strategies accordingly, as in the design of My Secret Pal in the From the Dark Side sidebar.

Persuasive technologies also take advantage of information about their target users. They can collect this information themselves or glean it from other sources, such as the Internet. Suppose our proponent of fruits and vegetables learned from a friend that the stranger he was trying to persuade suffered from a chronic iron deficiency. He could then leverage this information by connecting the consumption of spinach to the stranger's need for iron. Likewise, a persuasive computer game might access online medical records, learn about the player's deficiency, and automatically adjust its plot line, so an interactive spinach "character"—even Popeye himself—became the hero and wielded an iron sword in battle.

We propose two principles for the design of persuasive technologies with regard to the collection and manipulation of information about users. The first is: The creators of a persuasive technology must ensure it regards the privacy of users with at least as much respect as they regard their own privacy.[3]

To complement this principle, we should consider whether personal information is shared with a third party or used exclusively by a particular technology—whether the persuasive technology plays the part of big brother or little sister. So, consider a technology that monitors a child's water usage at the sink. It might chastise children who leave the water running while brushing their teeth. Here, information collected about personal water habits goes to an immediate persuasive end. We term this a "little sister" technology. What if a similar sink kept track of whether restaurant employees washed their hands, so their employer could later reward or punish them appropriately? Here, the same kind of private information—water use in the bathroom—is collected and disseminated to a third party. We term this a "big brother" technology.

Technologies providing us with information about ourselves are little sisters; technologies providing us with information about others, big brothers. Because they relate with users one-to-one, in a way that preserves the privacy of personal information, little-sister persuasive technologies are the less likely of the two to violate or infringe upon our privacy. Such violation leads us to our second design principle regarding user information: Persuasive technologies that relay personal information about a user to a third party must be closely scrutinized for privacy concerns.

[3]Granted, this principle (like our others) does not guarantee ethical technologies. Some designers with limited regard for their own privacy might not respect the privacy needs of all users adequately.

The Disclosure Principle

Some persuasive methods depend on persuaded parties not realizing they are being persuaded—or, more often, not realizing how they are being persuaded. We are far less likely to believe a car salesperson regarding the quality of a used car than we are to believe testimonials by people with no stake in persuading us to buy it. Knowledge of the presence of persuasive mechanisms in a technology may sensitize users to them and decrease their efficacy. Therefore, in some cases, such knowledge might diminish the effectiveness of a generally positive persuasion. This reasoning led us to our design principle: The creators of a persuasive technology should disclose their motivations, methods, and intended outcomes, except when such disclosure would significantly undermine an otherwise ethical goal.

For example, most simulations require technology to be more than just an exercise in imagination and role-playing. With the help of computers, people can experience a world not quite like their own, in order to be persuaded to change actual attitudes and behaviors. We distinguish between two kinds of simulation—integrated and immersive. An artificial infant intended to persuade teenagers not to become teen parents typifies integrated simulation. It brings a simulated baby into an otherwise real world; when the simulated baby cries, the cry is heard by real people in real situations. By contrast, an immersive simulation is one in which individuals take part in a fully virtual world, as in flight simulators and multiuser domains.

Because immersive simulations are by nature rich circumstantial experiences, full of cause-and-effect relationships, and because integrated simulations can interact realistically with outside variables, the creators of both must be sure to anticipate unexpected outcomes. For instance, the simulated baby might cry the night before a student's SAT test, hurting his or her performance and subsequent college admissions options. Or a student might find he or she enjoys carrying around an infant, even if it cries now and then, and be encouraged to reproduce right away. These are certainly reasonably predictable outcomes and must be addressed by the designer—perhaps with an emergency shut-off switch on the infant or with teacher oversight to confirm it conveys the correct message.

The creators of persuasive technologies, and especially simulations, must hold themselves responsible for all reasonably predictable outcomes of their persuasive methods. Such reasonable prediction requires significant user testing and holistic forward thinking on the part of designers.

The Accuracy Principle

Persuaders often tweak the truth or ignore it entirely. Because we expect this sort of behavior, we often regard someone trying to persuade us to do something, say, buy a car, with suspicion. We wonder what they are leaving out, then try to check what they tell us against more reputable sources. Our instincts also help us notice signs of dishonesty, from how much a person is sweating to the vibration in the person's voice. A good liar must show more than just a poker face, but a poker body too.

Computers can seem to lie with equanimity, so the user can't distinguish between false and true information. And people tend to trust the information computers deliver to them (see Tseng et al.'s Credibility and Computing Technology" in this issue). We have no reason to believe that a device monitoring our heart rate will deliberately misreport it. But imagine a scale meant to encourage healthier eating habits. It might be programmed to tell a teenage girl she weighs less than she actually does to minimize the chance of her developing an eating disorder. The motivation and the intended outcome of this persuasion are positive, but by mis-

Pack-a-Day Parent

Pack-a-Day Parent

The Golden Principle is not without limitations. A father who happens to be a pack-a-day smoker might design a technology to persuade his teenage son to stop smoking even if he, for whatever reason, would not consent to being so persuaded. But the father's contradictory motivations appear to conflict with the Golden Principle.

This conflict need not condemn the father as an unethical designer, especially when such factors as addiction are at play. While the violation of any one principle requires that the design be scrutinized more carefully, it does not necessarily jeopardize the ultimate ethical nature of the technology that was created.

reporting information, the technology risks being contradicted and thus devalued as a persuasive agent. The user might subsequently mistrust all persuasive technologies.

Therefore, established computer credibility is valuable—for persuasive purposes and for many other applications in society. Most humans anticipate dishonesty in other humans, sensing it to varying degrees. They do not, however, expect dishonesty from technology, nor do they have any instinctive aptitude for detecting it. To safeguard this credibility, and avoid its abuse, we therefore propose another principle for the design of persuasive technology: These technologies must not misinform in order to achieve their persuasive ends.

The Golden Principle

To round out these guidelines, we need to postulate a final "golden rule" of persuasion: The creators of a persuasive technology should never seek to persuade anyone of something they themselves would not consent to be persuaded of. We find support for this golden rule in the work of the 20th-century philosopher John Rawls, a Harvard professor who proposed in his landmark 1989 book *A Theory of Justice* that we consider ethics from behind a "veil of ignorance" [5]. Imagine you had no idea who you were in society, whether you were rich or poor, this or that ethnicity, male or female. Rawls contended that you would agree to obey only the ethical rules that benefited you no matter who you turned out to be. Similarly, if you imagined creating guidelines for an act of persuasion without knowing whether you were the persuader or the person being persuaded, you would want to make sure the persuasion would benefit both sides—in case you turned out to be the person being persuaded (see the sidebar Pack-a-Day Parent).

Some people might want to persuade others and consent to being persuaded of things that many find objectionable, say, to abort a fetus. However, when tempered with our other principles, the golden rule principle minimizes the potential for the ethical abuse of persuasive technology. For example, since abortion is already an ethically controversial act, any persuasive technology designed in light of our first principle—on outcomes—would already have had to wrestle with inherent and problematic issues.

We also should note the cultural context in which persuasion takes place. When stating anything about the ethical nature of a technology's design, not to qualify its appropriateness to a particular culture is to speak loosely. For instance, a persuasive doll intended to reduce teen pregnancy might be embraced as ethical in the U.S. but not in a culture that values early marriage and frequent childbirth. The creators of persuasive technologies that might go beyond the bounds of their own cultural systems should therefore be attentive to reasonably predictable ethical issues associated with their transfer into other cultural systems. While full treatment of how cultural differences influence the practice and perceived ethics of persuasion is beyond our scope here, further attention to this issue is valid and vital.

Conclusion

Our intent here is to persuade you to think critically about ethical issues at the convergence of technology and persuasion. However, remember that to analyze the motivation behind a persuasive act, it is important to put aside for a moment the intended outcome and ask, Why intend that outcome in the first place? But why should we want to persuade you? Because in the near future, persuasive technologies will be commonplace, affecting many people in many ways. By initiating this dialogue in the professional community and by proposing a first set of principles for persuasive design efforts, we hope to steer the field in a positive direction from the outset.

Our method is mostly an appeal to rational thinking, using the medium of this magazine to amplify our message. The outcome is, of course, in your hands. C

References

1. Andersen, K. *Persuasion Theory and Practice*. Allyn and Bacon, Boston, 1971.
2. Edgar, S. *Morality and Machines.* Jones and Bartlett, Sudbury, Mass., 1996.
3. Fogg, B., and Nass, C. Silicon sycophants: The effects of computers that flatter. *Int. J. Hum.-Comput. Stud. 46* (1997), 551–561.
4. Friedman, B. Human agency and responsibility: Implications for computer system design. In *Human Values and the Design of Computer Technology*. Cambridge University Press, Cambridge, Mass., 1997.
5. Rawls, J. *A Theory of Justice*. Harvard University Press, Cambridge, Mass., 1989.
6. Tobias, A. *The Only Investment Guide You'll Ever Need*. Harcourt Brace, Orlando, Fla., 1996.

Daniel Berdichevsky (dan@demidec.com) is executive director of DemiDec Resources, an educational firm based in Los Angeles, and an associate manager of the Persuasive Technology Lab at Stanford University's Center for the Study of Language and Information in Palo Alto, Calif.

Erik Neuenschwander (erikn@leland.stanford.edu) is a researcher in the Persuasive Technology Lab at Stanford University's Center for the Study of Language and Information in Palo Alto, Calif., and a computer consultant in northern California.