AI Designed With Humans in Mind

ANGELIQUE TAYLOR

In 2018 and 2019, two Boeing 737 MAX airplanes crashed, killing a total of 346 people. Accident investigators believe that in both cases, one of the main causes was an autonomous software system intended to prevent stalls. The system initiated actions for the aircraft without input from the pilots, who may have been unaware of the system's existence. As the pilots attempted to pull their aircraft up, the autonomous system, which was receiving information from a faulty sensor, continuously pushed the nose of the aircraft down. Later investigations into the two crashes highlighted a key design flaw: the pilots did not know how to override the artificial intelligence (AI) software controlling the aircraft when that system failed or malfunctioned. In the aftermath, experts called for designs that allow human pilots to intervene and provide better oversight for AI aircraft systems.

The Boeing crashes revealed the risks of highly automated systems in real-world environments. Among other lessons, they point to the need for systems that assign responsibility when AI fails. And, building on that, human operators need to be aware of how the AI works and how to perform essential tasks should the need arise. And of course, engineers also need to consider what to do if a computer system is hit by a malicious attack.

In his new book, *Human-Centered AI*, Ben Shneiderman examines the myriad risks that arise from the increasingly important role AI plays in human activities. But his focus is more on how AI presents opportunities for augmenting human abilities and meeting societal needs. Shneiderman, an emeritus

distinguished professor at the University of Maryland in the computer science department, has long conducted research on human-computer interaction, user interface design, information visualization, and social media. He brings these elements to bear on his discussion of AI, making the book especially relevant to AI researchers and developers.

Shneiderman introduces the human-centered artificial intelligence (HCAI) framework, which offers a vision for AI researchers, developers, business leaders, and policymakers to rethink the design of autonomous systems. In contrast to a technology-focused or data-driven approach that focuses on algorithm performance metrics such as speed or accuracy, human-centered thinking stems from the perspective of the system's human users. In HCAI, the right initial question is "How useful and interpretable is the system for users?" rather than "What is the system capable of?"

The HCAI framework includes "combined designs" of AI-based algorithms and human-centered thinking that "amplify, augment, empower, and enhance people," in Shneiderman's words, enabling people to do their work more fluently. He recommends using what he calls supertools—telebots, active applications, and control centers, for example—instead of robots or other automated physical systems. HCAI views AI technology as a tool to empower users, not replace them.

At its core, human-centered design forefronts the need to maintain human control over AI to ensure that these systems are safe, reliable, and trustworthy. Shneiderman recommends continued testing and evaluation of AI with users, with ongoing oversight to ensure performance metrics are met through retrospective analysis. Understanding how the AI has succeeded or failed over time can prepare designers to plan for future situations users might encounter.

Human-Centered AI
by Ben Shneiderman. New York, NY: Oxford University Press, 2022, 400 pp.

Shneiderman illustrates this point with examples of successes and failures of recent technologies that are highly automated but subject to human control. Among the successes he includes smart household thermostats, which allow users to remotely monitor and control their energy usage—sometimes from their phones. He also describes potential failures such as an autonomous vehicle that, without the ability for a human to take the wheel, could leave a passenger stranded—or worse. These are useful, real-world instances of how the HCAI framework could work in practice and may motivate developers and researchers to adopt it.

Among the goals of HCAI is the production of systems that are predictable, secure, fair, and controllable. But this last aspect depends heavily on the context in which the technology is used. For example, manufacturing may require high automation and low human control for tasks that are rapid and repetitive. On the other hand, tasks requiring a high degree of human mastery, such as the kind of minimally invasive surgery for prostatectomies and other procedures that surgeons can perform with the robotic Da Vinci Surgical System, are well suited for AI systems with high human control and low automation.

According to Shneiderman, employing the principles of HCAI can also help avoid costly business mistakes. Many robotics and AI companies have failed because their products weren't useful. In particular, he points out that many human-robot interaction (HRI) researchers start from the assumption that given the right morphology (appearance), size, and interaction profile, robots can function as social actors and successfully interact with people. This has turned out not to be the case; humanoid robots such as Honda's Asimo, though a hit at trade shows, often have difficulty performing even basic practical tasks. Shneiderman recommends that HRI researchers glean inspiration from the success of widely adopted simple tools, such

as smart thermostats, robot vacuums, or Google documents, which make humans' tasks easier without resembling humans at all.

Going beyond these simple tools, as AI becomes integrated into highly complex and essential systems such as air travel, HCAI requires that an organization manage its AI systems in ways that make them transparent and accountable. Following the practice of civil aviation, Shneiderman suggests that AI systems should include audit trails to record data used by the system—like the "black boxes" in airplanes, which are critical to investigations when things go wrong. These records can serve as evidence to assign legal liability for mistakes and exonerate the falsely accused. Enabling such retrospective analysis of adverse events could help prevent future occurrences, and he cites examples of successful audit trails documenting incidents that occurred in systems as diverse as aircraft, the stock market, industrial robots, and autonomous vehicles.

Some related open questions that Schneiderman addresses in the book include: What data is required for retrospective analysis? How can researchers analyze large datasets? Who should have access to data while preserving privacy and avoiding falsification? Although these questions of responsibility and forethought are key to whether AI serves society's needs, public discussion of AI has largely been shaped by older cultural narratives.

For example, misrepresentations of AI in the media have greatly impacted public perceptions of the technology. Flashy headlines ("Robots Can Now Read Better Than Humans," "How Robot Hands Are Evolving to Do What Ours Can," "How DALL-E Could Power a Creative Revolution," etc.) often overstate the ability or utility of the technology. More realistic and thoughtful media representation of robots and AI—including depictions in films, TV, and fiction—could provide people with a better sense of the technology and what it's capable of. What's more, including users in the development of intelligent systems could help mitigate concerns of widespread unemployment and other negative effects of AI.

In reading this book, I appreciated many of Shneiderman's ideas about human-centered design that align with my own research on human-centered algorithm design for healthcare environments. My work incorporates a collaborative process to design robotic systems for use in hospitals—much like the human-centered approach discussed in the book. Our design process involves interviewing potential users, including nurses and doctors; understanding how to integrate a new system into an environment it has not been used in before; and reflecting on how such a system might affect the workload of the healthcare team.

My research experience also leads to a point of disagreement with Shneiderman's assertion that people should be using supertools instead of robots. He paints a picture of computers as tools only, instead of potential teammates, partners, or collaborators. His ideal of a supertool is a computer that facilitates human-human teamwork by providing what he calls "superhuman perceptual and motor support."

But some situations require what roboticists call "embodiment"; sometimes robots can do things physically that humans simply cannot. The Da Vinci system, for example, can use smaller incisions and cameras that allow it to accomplish tasks in areas that are difficult or impossible for a human surgeon using traditional techniques. Similarly, perhaps Shneiderman's HCAI principles can be applied to intelligent systems that require such physical embodiment, but with user interfaces that allow people to understand and interpret intelligent system decisionmaking—leading to more useful human-robot interfaces.

A challenge of the HCAI approach is that it requires many resources, including software engineers, designers, policymakers, and business leaders. Not all organizations have these resources on hand when trying to develop useful AI systems, although of course it depends on the context; some well-equipped organizations may be able to build the many relationships that are required to execute the HCAI approach. HCAI also demands time for planning and relationship-building among the very different groups involved, among other critical tasks. And time can be a scarce resource, especially in the private sector.

Shneiderman's HCAI framework promotes participatory design, involving users in the development of useful AI. I wish he provided more discussion on how readers can deploy this design method beyond conducting user interviews. In my experience, there are many techniques beyond participatory design that could be used to explore and enhance AI design. These include storyboarding, using design fiction, and futureproofing, which is a planning process intended to minimize the negative effects of future events. Expanding the variety of inputs into AI design will be essential to achieving Shneiderman's transformative vision of a more human—and humane—future.

Angelique Taylor *is a professor at Cornell Tech, a campus of Cornell University in New York City, focusing on robotics, healthcare, and AI that improves human teams in real-world environments.*