BEN SHNEIDERMAN

Human-Centered AI

Computer scientists should build devices to enhance and empower—not replace—humans.

In early 1997, before Alexa and Siri were conceived, I engaged in several public debates with MIT Media Lab professor Pattie Maes about the future of computing and artificial intelligence. The starting place for our disagreement was the question of control. Maes was advancing a vision of machine autonomy, of the proactive "software agent" that "can take initiative because it knows what your interests are." I offered a different vision: software designs that give users high levels of understanding and control over their AI-enabled devices to preserve human agency.

Although both views of AI design have their value, I believe that pursuing a vision of AI and technology design that starts with machine autonomy is dangerous for the future of human beings, our societies, and our environment. The consequences of an autonomy-first approach have begun to make themselves felt, in stock market flash crashes, deadly failures of autonomous missile systems in the second Iraq War, fatal accidents involving self-driving cars, and the two Boeing 737 MAX crashes. Autonomy-first elevates the dangers of hidden biases in algorithms for making decisions about who gets a mortgage, who gets a job, who gets paroled, and in the "surveillance capitalism" that turns surreptitiously collected private data into targeted advertising.

As AI-enabled devices proliferate in the economy and our daily lives, we are at a decisive moment. What will it take to put humanity on a path where humans steer AI to advance their own values and aspirations, rather than accepting control by autonomous decision algorithms that remain beyond our view and direct influence?

We should reject the idea that autonomous machines can exceed or replace any meaningful notion of human intelligence, creativity, and responsibility. The clichéd images of a human hand shaking a robot hand or a humanoid robot pretending to think are archaic and misguided. People have remarkable abilities, and in contrast to future visions of humanoid robots and driverless cars that relieve us of our roles and responsibilities in daily tasks, I look forward to next-generation technologies that bring greater human control of ever-increasing automation.

My agenda is to help change the way we imagine, talk about, and design AI systems, starting with a vision of people working together, in charge of technology, and benefitting from information-abundant displays that enable them to ask better questions and make bolder decisions. A growing community of AI researchers are calling this approach Human-Centered AI (HCAI). Our goal is to amplify, rather than erode, human agency.

And more. HCAI is a vision of how machines might augment humans, and even encourage our best impulses toward each other, rather than how they might replace

humans with something supposedly better. To understand HCAI's promise not only for our machines but for our lives, a good starting place is an appreciation of the two competing philosophies that have shaped the development of AI, and what those imply for the design of new technologies. For policymakers, comprehending these competing imperatives can provide a foundation for navigating the vast thicket of ethical dilemmas now arising in the machine-learning space.

Aristotle vs. da Vinci

The difference between artificial intelligence and Human-Centered AI reflects a historical and philosophical clash between two approaches to gaining knowledge about the world—Aristotle's rationalism and Leonardo da Vinci's empiricism.

Rationalists believe in logical thinking, which can be accomplished in the comfort and familiarity of your lab or even your living room. They believe in the perfectibility of rules, the strength of formal methods, and the constancy of well-defined boundaries between different categories of things—such as hot and cold or wet and dry. Aristotle recognized important distinctions, such as the differences between vertebrates and invertebrates, and the four classes of matter: earth, water, air, and fire.

As useful as strong rules and categories are for allowing clarity of thought, they can limit seeing other options and middle grounds. Followers of rationalism have included Descartes, with his powerful assertion of the duality of mind and body, and his mechanistic view of nature, and, in the twentieth century, the famed statistician Ronald Fisher, whose overly rigid commitment to statistics led him to reject early data on the risks of smoking. Rationalism, especially as embodied in logical mathematical thinking, is the basis for much of AI research, in which algorithms are treasured for their elegance and measured by their efficiency. This was true in the days of symbolic AI algorithms written by programmers, and remains true in our time of data-driven deep learning based on neural networks.

By contrast, empiricists believe that researchers are enriched by the contradictions and ambiguities that come with real-world experiences in all their contextual complexity and diversity. They understand that beliefs have to be continuously refined to respond to changing realities and new contexts. Da Vinci developed fluid dynamics principles by using his keen eyesight to study bird flight and watch water flowing around obstacles. Galileo was following da Vinci's method by noticing the rhythmic swaying of a chandelier in church, which led him to the formula for pendulum swing times. David Hume, who recognized that correlations between events were fallible predictors of future consequences, was another great empiricist. The statistician John Tukey, in contrast to Ronald Fisher, believed in looking at data graphically, because analysts could see errors, missing data, anomalies, exceptional performances, and unexpected distributions of data. This graphical view revealed surprising patterns and suggested fresh questions.

Empiricism, pursued through the empathic observation of people, is the basis for much of the HCAI community's work, which assesses human performance so as to improve it. Empiricists question simple dichotomies and complex hierarchies, because these may sometimes limit thinking, and undermine the analysts' capacity to see important nuances and nonhierarchical relationships.

The rationalist viewpoint, however, is dominant in the AI community. It leads researchers and developers to emphasize data-driven solutions based on algorithms. Rationalism also favors the belief that statistical methods and machine learning algorithms are sufficient to achieve AI's promise of matching or exceeding human intelligence on well-defined tasks, rather than requiring deep engagement with domain experts who understand causal relationships among variables. Some AI advocates have gone so far as to say that theories about causal relationships are no longer needed for AI because the amount of data now available for machine learning allows for correlations that are strong enough to guide decisions.

Such reasoning also leads to the conclusion that machine learning can replace expert knowledge. And though it is true that machine learning can reveal patterns in the data used to "train" algorithms, it does not deal with surprising extreme cases, such as when Tesla's self-driving car could not distinguish a white truck from the sky and drove right into it. Learning from these and hundreds of other incidents will lead to safer cars sooner. But standard approaches to machine learning do not allow adequate learning from such apparently exceptional events. Efforts to develop commonsense reasoning, explainable AI, and causal understanding seem still shallow compared with what humans do when they formulate problems, find innovative solutions, and—as I am doing here—raise challenges to existing beliefs.

Human curiosity and desire to understand the world mean that humans are devoted to causal explanations, even when there is a complex set of distant and proximate causes for events. Machines lack such curiosity. Algorithms trained via machine learning are still subject to failure in novel situations, because they lack the innately human attributes of common sense and higher cognition. For example, machine learning may not recognize hidden biases or the absence of expected patterns.

And though humans experience the passage of time and adjust to new realities, computers with machine learning repeat the past, including its mistakes. The Google Flu

Trends project, which began in 2008, demonstrates this problem. It aimed to predict flu outbreaks by studying patterns of internet search queries for items such as tissues and flu medications. It worked for a while, but over time, as people used search engines in new ways, and the underlying search algorithms were revised in response, these changes led to prediction failures that were embarrassing enough for Google to shut down the website and project.

Human learning

My own professional journey began when, as an undergraduate math and physics major, I wrote enthusiastically about the possibilities of artificial intelligence from a rationalist perspective. But after graduating, when I began to teach data processing at a community college, my views began to shift. Close contact with students studying programming and information systems gave me insight into how diverse people faced varied challenges in learning something that seemed obvious to me. If I was to help them learn, I had to appreciate their struggles.

My PhD research focused on optimization of data structures for the emerging topic of database systems. I was still building on the rationalist tradition, using quantitative methods that appealed to my sense of rigor. Then as an assistant professor in computer science at Indiana University, and to the surprise of my computer science colleagues, I partnered with a young psychologist, Richard Mayer, who trained me in experimental methods and statistics for research on human subjects. I began to study human performance in laboratory-like settings, to improve design of programming languages and tools. I moved on to complement these controlled experiments with qualitative case studies, and a broader focus on research on designing technologies to empower people—just as personal computers were emerging. I was able to experience the satisfaction of contributing to the development of a new field: human-computer interaction.

In adopting this empiricist approach, I had come to see that engaging with users of technology could lead me to fresh insights. I used (and continue to use) naturalistic observations, usability studies, and repeated weeks-long case studies with users doing their work to complement the rationalist approach of controlled experiments in laboratory settings. This work was aligned with the rise of the concept of design thinking, an approach to innovation that begins with empathy for users and pushes forward with humility about the limits of machines and people. Empathy enables designers to be sensitive to the confusion and frustration that users might have and the dangers to people when AI systems fail. Humility leads designers to recognize the inevitability of failure and inspires them to be always on the lookout for what wrongs are preventable.

Humans in the group; computers in the loop

The AI community's sympathy for rationalism continues to lead developers to favor autonomous designs in which computers operate reliably without human oversight. This approach typically brings with it an imagined future of intelligent machines that pretend to have emotions. Many developers believe that autonomous machines will seem less threatening to us if they act like friends or partners. But computers don't have emotions; people do. Today, humanlike social robots remain novelties, mostly confined to entertainment, with only voice user interfaces succeeding as a consumer product.

HCAI designers recognize that humans are happily and productively woven into social networks; for example, at work we are embedded in social structures of supervisors, peers, and staff who we want to please, inspire, and respect. From the HCAI perspective, computers should play a supportive role, amplifying people's ability to work in masterful or extraordinary ways.

Although a growing number of people are demanding that AI machines include a "human in the loop," this phrase often implies a grudging acceptance of human control panels. Those who seek a complete and perfect system are resistant to the idea that there needs to be human intervention, oversight, and control. The HCAI bumper sticker would be *Humans in the group; computers in the loop*.

I believe that progress in technology design will accelerate as recognition spreads that humans must have meaningful control of technology and are responsible for the outcomes of their actions. When humans depend on automation to get their work done, they must be able to anticipate what happens, because they, not the machines, are responsible. One effective way to enable users to anticipate the consequences of their actions is with direct manipulation designs—the objects and actions are represented on the screen; humans choose which actions to carry out; the actions and objects are all visible. The file drops into the trash can, you hear the clinking and know what's going on. Touch screen, swipe left and right. Overview first, zoom and filter, then details on demand. Humans are in control; computers are predictable.

Although AI projects are often focused on replacing humans, HCAI designers favor developing information-rich visualizations and explanations built in, rather than added on. Today, the vast majority of apps are giving users more control, not less—by showing highway navigation routes on maps, exercise histories in bar charts, and financial portfolios in line graphs. These displays give users a clear understanding of what is happening and what they can do. Visual displays are now frequently complemented by audio interfaces based on speech recognition and generation, opening up new possibilities for diverse users to accomplish their tasks. Digital cameras have lots of AI embedded, but the users get to take the picture they want—and more advanced cameras increase,

rather than decrease, the user's control over the device.

Similarly, educational intelligent tutoring systems with chatty humanlike avatars have given way to widely used online courses where users chart their own progress, see the gaps in their learning, and click to decide how fast they will move forward. In my view, future technologies are more likely to be what I call supertools and active appliances, rather than teammates, partners, and collaborators. Tele-operated drones, home controls, and surgical devices will spread, and the ambitious control rooms for NASA's Mars rovers, transportation management centers, patient-monitoring displays, and financial management software, such as Bloomberg Terminals, will become the compelling prototypes for still more advanced applications.

Accountable AI

This human-centered empiricist-driven strategy should apply strongly to military applications where responsibility within a chain of command is a core value. Even when the case for autonomy in defensive systems is strong, no weapons systems should be fully autonomous. But the HCAI approach also applies to the popular notion of autonomous vehicles or self-driving cars, where attaining adequate levels of safety will require an empiricist's outlook and design of effective user interfaces to enable meaningful human control, even as the levels of automation increase. *Self-driving* should become *safety-first* cars in which proven methods such as collision avoidance are improved by better user interfaces. Then further improvements will come from vehicle-to-vehicle communication, improved highway construction, and advanced highway management centers that build on the strategies of air traffic control centers.

Devotees of autonomous design often assume machines will do the right thing. But as flaws such as algorithmic bias in AI-driven systems emerged to shatter the belief in the perfectibility of these systems, HCAI researchers have begun to take on fairness, accountability, transparency, explainability, and other design features that give human developers, managers, users, and lawyers a better understanding of what is happening than in the previously closed black boxes. Moreover, many types of AI systems should include logging activity to support transparent and retrospective review of failures and aggregate patterns of usage. This strategy, following civil aviation's commitment to continual improvement, would install the equivalent of a flight data recorder in every robot. Adding audit trails or activity logs would assure appropriate accountability, especially in applications that have significant consequences for people and organizations, in settings ranging from the hospital to the battlefield. The design goal is to enable retrospective analyses of failures and near misses and review of aggregate patterns of performance to allow continual design improvement.

In sum, the HCAI community's sympathy for empiricism leads its members to design systems with users at the center of attention. To get a sense of how HCAI can enrich human experience in ways that conventional AI cannot, consider its application to the daily life of an aging population. Conventional AI might lead toward home robots that clear dinner tables to fill traditional dishwashers. I see this as eroding human agency. AI isn't the answer to everything: why not have small dishwashers built into, under, or beside dinner tables so that an older person could simply slide plates, cups, and cutlery into the dishwasher, which washes them to make them readily available for the next meal. This approach enhances self-efficacy rather than waiting for a mobile robot to carry off the dishes to the kitchen and bring them back when requested.

In contrast, Human-Centered AI could promote community-based solutions for some problems. For older adults and people living with disabilities, the Meals on Wheels programs that deliver prepared meals are much appreciated, not only for the food, but for the social contact with the delivery person, who may bring the food in, check on the recipient, and maybe clear the table. Those who deliver the meals gain satisfaction in helping those in need and might be celebrated in their communities. AI algorithms could help match interests and personalities of deliverers and recipients, while scheduling efficient routes.

We should reject the idea that autonomous machines can exceed or replace any meaningful notion of human intelligence, creativity, and responsibility.

Social support for professional caregivers, family members, and helpful friends could increase the pride and satisfaction they get from taking compassionate care of homebound individuals. Smartphone applications could weave together networks of professional caregivers with income-earning strategies that give workers a better deal than existing gig-economy schemes. However, the special needs of older adults and people with disabilities would prioritize long-term commitments that build relationships, and opportunities for care recipients to rate the caregivers, possibly with public comments of praise that would raise status for caregivers. Rationalist AI treats caregivers, aging people, and people with disabilities as

passive targets of automation; HCAI views them as people whose lives can be enriched through a strengthened sense of agency and social connectedness.

HCAI for policymakers

Policymakers must now deal with the challenges of AI-driven decision processes across a vast range of applications, including mortgage and parole decisions, self-driving cars, and medical implants. Congressional committee and government agency staffers are considering how to regulate social media platforms to reduce fake news, hate speech, criminal activity, and terrorist recruitment. Government policymakers are working hard to understand and make decisions about how to deal with AI technologies that impact public services, such as whether to allow police to use facial recognition, how to limit foreign agents from interfering with voting, and whether to permit immigration agencies to screen refugees with opaque algorithms.

These policy concerns are usually viewed through an ethical lens, and have led to more than 400 reports about responsible, fair, transparent, and accountable AI. But building the edifice of reliable, safe, and trustworthy AI systems requires not just attention to ethics, but actual design decisions that embody the HCAI approach. Ethical principles and design guidelines from leading companies such as Apple, Google, IBM, and Microsoft are helpful, but insufficient to bridge the gap between ethics and practice. General principles such as "mitigate social biases" or "AI systems must be transparent and explainable" are fine, but specific guidelines, clear examples, and assessment criteria are needed for software engineers to know when they have done a good job.

I propose that design, development, and implementation of AI systems needs to adhere to these three HCAI guidelines:

1. Build reliable and transparent systems based on sound software engineering team practices, such as testing software to check AI algorithms, using visual tools to reveal anomalies, and testing databases to enhance fairness in machine learning training datasets. A big step forward would be to adapt the model of flight data recording, which has contributed so much to make civil aviation safe, to record activity that allows retrospective forensic analysis after failures and near misses. Another vital feature will be to support explainable AI, which enables users to understand AI-based decisions and seek redress for what they see as unfair or incorrect decisions. New approaches that involve visual user interfaces are proving to be helpful in giving designers, implementers, and users a better understanding of AI-based decisionmaking. A still better approach is to prevent the need for explanations by using visual control panels that let users understand their progress through the process.

2. Pursue safety culture through effective business management strategies, such as clear demonstrations by corporate leadership of their commitment to safety. Hiring risk-assessment professionals, allocating adequate budget for safety and training of staff, and reporting all failures and near misses are key pillars of safety culture. Another is adherence to voluntary industry standards and guidelines promoted by professional groups such as the Robotics Industry Association, Underwriters Laboratory, the International Standards Organization, and IEEE, the largest engineering society. HCAI thinking suggests incident reporting and suggestion box schemes, such as the Food and Drug Administration's Adverse Event Reporting System, the Federal Aviation Administration's Aviation Safety Reporting System, and the Partnership on AI's Incident Database, now with more than a thousand entries.

3. Increase trust through certification and independent oversight within each industry, carried out by government agencies, independent accounting firms, insurance companies, professional societies, and nongovernmental organizations. An excellent model is the National Transportation Safety Board, whose investigations of airplane, ship, train, and car accidents are widely respected, and whose recommended improvements are thus taken seriously by the industries the board oversees. Though a National Algorithms Safety Board is an appealing notion for independent oversight, a more realistic approach is for existing government agencies to increase their capacity to investigate consequential and life-critical failures in each industry.

Above all, HCAI offers fundamental philosophical and design principals for addressing the challenges created by the diffusion of AI systems through society and the economy: start with the users. For example, if Facebook and other social platforms won't give their users greater control of filters on what users get to see and what ads they are shown, then users could be assured such control through regulation or new legislation.

AI advocates might claim that AI regulation by government will slow innovation. But the history of automobile safety and efficiency requirements shows that they led to massive innovation and public benefits. Regulation of drug safety has similarly been essential to the commercial success of the pharmaceutical industry. Linking the development of AI systems to guidelines that assure transparency and accountability will enhance innovation, public confidence, societal value, and commercial value.

Government research agencies can also lead the way by incentivizing HCAI approaches to cutting-edge

Linking the development of AI systems to guidelines that assure transparency and accountability will enhance innovation, public confidence, societal value, and commercial value.

research. For example, the National Science Foundation's program for National AI Research Institutes seeks to accelerate "research, transforming society, and growing the American workforce." This program's request for proposals lists "Human-AI Interaction and Collaboration" as the first of eight research themes. The program description supports human-centered design by encouraging attention to "ethics, fairness, privacy, lack of deception, explainability, protection of vulnerable and protected populations, and participatory and inclusive design," so as to produce "trustworthy and safe" AI systems.

But achieving these laudatory goals demands more than good intentions; it also requires institutional mechanisms that facilitate HCAI. Such mechanisms are compatible with the NSF program's requirement that investigators address foundational AI research while conducting "use-inspired research" that "drives innovations in related sectors of science and engineering, segments of the economy, or societal needs" while working with "external stakeholders such as industrial partners, public policy-makers, or international organizations." NSF should link these mechanisms with the societal goals of its AI program by preferentially funding projects that take human-centered design seriously.

The encouragement for AI researchers to raise their emphasis on use-inspired research is aligned with the growing awareness of the benefits of partnering with nonacademic organizations and researchers. When AI researchers partner with those who are close to real-world problems, the quality of foundational research is likely to increase, as are the benefits to society. Such persistent interactions enhance the mutual understanding—empathy—that underlies the HCAI approach. It also brings researchers closer to users, the other pillar of HCAI's commitment to empiricism.

These efforts are a good start, but much more needs to be done in educating AI researchers and practitioners about the philosophy, principles, and methods at the core of the HCAI framework. Making sure that students interested in AI receive a diverse education that includes humanities and social sciences will also help. When the traditions of rationalism and empiricism are refined to account for the richness of modern technologies and the distinctions between machines and people, then I believe we are more likely to see an increased capacity to create reliable, safe, and trustworthy technologies.

AI can advance in ways that support, rather than erode, fundamental human aspirations. For many people, the desire for social connections with trusted friends, family, colleagues, and neighbors is strong. People seek to be masters of their own lives, strive to express their creative potentials, and with some exceptions, take responsibility for their actions. People are imperfect and sometimes malicious, but I want to support technologies that help people to realize their hopes, help others when they can, and contribute to causes they believe in. I envision a future that emphasizes human self-efficacy, creativity, responsibility, and community, while promoting better education, business, government, health and well-being, and economic development. I envision a future in which technology innovations respect human values, rights, and dignity, while promoting fairness, equity, and justice. HCAI can illuminate the technological pathways toward these goals.

Ben Shneiderman *is a distinguished university professor of computer science at the University of Maryland and a member of the National Academy of Engineering.*

RECOMMENDED READING

John Markoff, *Machines of Loving Grace: The Quest for Common Ground between Humans and Robots* (New York, NY: HarperCollins, 2016).

Robin R. Murphy, *Disaster Robotics* (Cambridge, MA: MIT Press, 2014).

Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York, NY: Broadway Books, 2016).

Ben Shneiderman, "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," *International Journal of Human-Computer Interaction* 36, no. 6 (2020): 495–504.

———, "Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems," *ACM Transactions on Interactive Intelligent Systems* 10, no. 4 (2020).

Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2019).

Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York, NY: Oxford University Press, 2016).

Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (W. H. Freeman & Co., 1976).