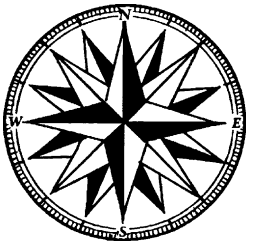

Book Review

Human-Centered AI, by Ben Shneiderman. Oxford: Oxford University Press, 2022. 305 pp.

Jay Killoran, Queen's University, Canada
Andrew Park, University of Victoria, Canada

Ben Shneiderman presents a compelling argument for why the future of artificial intelligence (AI) design should be human-centric. He begins with the tragic backstories of two Boeing 737 MAX airplane crashes in 2018 and 2019 that were largely due to algorithmic error. Investigations revealed that the cause of both crashes was the installation of autonomous software meant to prevent stalls, which kept pointing the nose of the planes downwards. Disturbingly, the developers of the autonomous control system believed that it was so reliable that pilots were not even informed of its presence, and thus they had no knowledge of what to do to retake control.

This example illustrates the danger of prioritizing computer automation over human control. There has been a long-held belief, admittedly even by Shneiderman himself, that by granting greater autonomy to machines there must be a corresponding decline in human control. Sometimes this is desired—advances in vehicle airbags keep people safe by deploying in a fraction of a second, far quicker than any human's reaction time. However, Shneiderman shows that excessive automation can lead to disaster. As another example, in 2016, one of Tesla's self-driving vehicles failed to distinguish between a white vehicle and the sky, resulting in a fatal head-on collision. Unfortunately, this is not an isolated event. As of April 2024,[1] there have been 44 Tesla autopilot deaths and hundreds of collisions.

The consequences associated with the pursuit of automation drove Shneiderman to change his perspective. Instead of viewing human control and computer automation as extremes along the same axis, he posits that reliable, safe, and trustworthy AI systems are designed with high automation *and* high human control. Shneiderman proposes a human-centered AI (HCAI) framework where he writes, "the goal is not to replace people but to empower them by making design choices that give humans

[1] Source: Tesladeaths.com.

Business Ethics Quarterly 34:3 (July 2024), pp. 517–521. DOI:10.1017/beq.2024.10
Published by Cambridge University Press on behalf of the Society for Business Ethics.
© The Author(s), 2024.

https://doi.org/10.1017/beq.2024.10 Published online by Cambridge University Press

control over technology." By prioritizing human control, Shneiderman aims to reduce the harms of AI failures, calm fears that robots will lead to mass unemployment, and give humans a sense of agency and accomplishment.

The book is structured into five parts. In part 1, Shneiderman provides an explanation of the harms we see from AI by mapping rationalism to traditional AI design. According to Shneiderman, rationalism has been the predominant philosophical basis for much of AI research and development, in which AI models are built in a lab and then deployed in real-life settings. While a rationalist approach offers a useful starting point, Shneiderman argues that it does not capture the full complexities of the physical world and may explain why catastrophes occur. This is consistent with business ethics scholarship on AI hiring (Bhargava and Assadi 2024) and algorithmic recommendations (Kim and Routledge 2022). In hiring, statistical algorithms are often more effective than humans at predicting the performance of job candidates, but they are less effective at predicting the compatibility of a candidate with colleagues. Similarly, algorithmic recommendations for assigning credit scores may be useful for integrating a wide range of parameters into an efficient score, but they have demonstrated a discriminatory bias against women.

Part 2 introduces the need for researchers and developers to take a human-centered approach to AI design and presents the HCAI framework, which is organized along two dimensions of high and low computer automation, and high and low human control. Many AI developers favor designs where AI operates reliably without human oversight, such as autonomous vehicles (AVs). This idea is compelling, but modern AVs are still imperfect and responsible for many collisions and fatalities, prompting business ethicists to posit design trade-offs and human control in their development (Scharding 2021). Shneiderman emphasizes that achieving reliable, safe, and trustworthy AI requires designing systems that have high levels of both automation and human control because it keeps users and designers responsible for preventing harms if AI malfunctions. Instead of pursuing AVs that operate without humans, he suggests orienting toward "safety-first cars" that use AI to improve parking assist, lane following, and collision avoidance.

Part 3 defines the two goals of HCAI research: the *science* goal aims to build systems that perform tasks to the same extent, or better, than humans, while the *innovation* goal aims to design systems that amplify human abilities to better perform work themselves. Shneiderman clarifies that some tasks are better suited to a science goal with greater automation—such as the use of AI programs for identifying breast cancer tumors—while other tasks are better suited to an innovation goal that emphasizes greater user control, such as surgical robots that support physicians during medical procedures.

Part 4 discusses the role of governance structures in bridging the gap between the ethical principles of HCAI and the practical steps needed to achieve them. Four levels of governance are discussed: software engineering teams that develop and implement AI, organizations that manage AI in a culture of safety, industry-specific oversight that provides trustworthy certification, and regulation by government agencies. Each level plays a vital role in ensuring that AI is humane and does not compromise human rights, dignity, privacy, or accountability. For instance,

https://doi.org/10.1017/beq.2024.10 Published online by Cambridge University Press

software development teams can implement audit trails to track and assign responsibility when undesirable outcomes occur; managers can form committees that review product failures; third-party auditors can support and approve HCAI projects; and governments can fund research that promotes the responsible use of AI.

Part 5 concludes with future directions for HCAI research and development, including techniques to adopt human-centric approaches to AI in academia, industry, and government, methods to measure the trustworthiness of HCAI systems, and strategies to support older adults through HCAI initiatives.

What is particularly refreshing is that Shneiderman does not take a utopian or dystopian view of the future of AI. Public sentiment commonly fluctuates between overly optimistic promises of AI and hyperbolic, negative depictions of AI replacing human jobs and threatening civilization. Media portrayals often exacerbate public sentiment with provoking headlines such as "Why the Godfather of A.I. Fears What He's Built" (Rothman 2023). Shneiderman takes a balanced approach to this conversation. While lauding the potential of AI to improve human productivity and well-being, he also offers a sobering account of the persistent risks that AI presents. He does this not to scare the reader; rather, he draws attention to these risks to highlight that harms from AI are not exceptional situations, but common realities. Shneiderman believes it is crucial to design human control into AI systems to prevent future harms from taking place.

There are several valuable takeaways from *Human-Centered AI*, two of which we wish to highlight. Regarding the previously discussed predilection toward rationalism in AI design, Schneiderman suggests a better approach. First, designers should begin with the theoretical foundation of rationalism and then test AI systems empirically to explore for contextual complexity and uncertainty. This would provide insight into the biases and extreme cases for which AI models have not accounted. Second, Shneiderman acknowledges that no amount of empirical testing will fully remove the risks posed by highly automated AI systems. This is why he underscores the need for designers to integrate human control alongside automation —the ability for humans to intervene will always be necessary.

The second strength is the thoroughness with which Shneiderman explains how designers can ensure that AI systems become reliable, safe, and trustworthy. Leveraging his background in computer science, he details how the design and testing phases of AI development can be performed in a manner that limits bias, enhances fairness, and maintains accountability. For instance, Shneiderman suggests that audit trails, which are commonly used in flight data recorders and are essential for understanding why aviation crashes occur, should similarly be implemented into AV, industrial robots, and stock market trading algorithms. He also recommends that software engineering workflows include user experience design so that users can understand how AI decisions are made and challenge them if they believe the logic is flawed. Moreover, he proposes explicit testing for known biases in past algorithmic performances, particularly racial, gender, and accessibility, which can then be mitigated before they are deployed. Finally, he identifies and recommends various explainable AI tools to reduce the opacity of modern AI

https://doi.org/10.1017/beq.2024.10 Published online by Cambridge University Press

systems. Rarely do business ethicists see such a meaningful integration of normative values with technically sound design principles.

One criticism of the book is that Shneiderman proposes a scale for measuring trustworthiness despite acknowledging the difficulty in establishing universal measurements of human-centered outcomes. The scale uses numerous subjective factors, such as "fairness was tested." We commend Shneiderman for his attempt to provide a measurement tool, but we question its usefulness in practice. The scale may unintentionally serve as a superficial tool for justifying more rationalistic approaches than for empirical testing and refinement.

Trustworthiness, dignity, and justice are inherently incommensurable. Perhaps the goal should not be to design explicit measurements of metaphysical, human-centered constructs. Rather, the goal should be to incentivize empirical testing of AI before deployment, and to introduce meaningful human controls in the face of rapidly evolving AI systems. When we abdicate control of AI systems, we create responsibility gaps that become difficult to reconcile (Bhargava and Assadi 2024). As Shneiderman emphasizes, the danger of highly automated systems is that accountability becomes muddled and error-prone systems go unchecked. Assessments are important, but quantitative standards should be reconsidered when evaluating ethical risks (Scharding 2021). We encourage business ethicists to reimagine incentives for AI designers to maintain human control and accountability.

REFERENCES

Bhargava, Vikram R., and Pooria Assadi. 2024. Hiring, Algorithms, and Choice: Why Interviews Still Matter. *Business Ethics Quarterly* 34 (2): 201–30.

Kim, Tae Wan, and Bryan R. Routledge. 2022. Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly* 32 (1): 75–102.

Rothman, Joshua. 2023. Why the Godfather of A.I. Fears What He's Built. *The New Yorker*, November 13. https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai.

Scharding, Tobey K. 2021. Recognize Everyone's Interests: An Algorithm for Ethical Decision-Making about Trade-Off Problems. *Business Ethics Quarterly* 31 (3): 450–73.

. . .

JAY KILLORAN (j.killoran@queensu.ca, corresponding author) is completing his PhD in management information systems at the Smith School of Business, Queen's University, ON, Canada. His research interests include human-computer interaction and technology ethics, with an emphasis on artificial intelligence and the future of work. He is currently studying moral responsibility, accountability, and blame attributions between humans and autonomous technologies in the workplace. His research also investigates how biometric technologies enable organizational efficiencies and productivity, while simultaneously constraining the dignity of employees and customers. He is passionate about preserving human well-being in the face of rapid technological change.

https://doi.org/10.1017/beq.2024.10 Published online by Cambridge University Press

Andrew Park is assistant professor of management information systems, University of Victoria, Victoria, BC, Canada. His research interests include information systems, and innovation and technology management. He is currently investigating how open innovation mechanisms impact value creation of firms within the emerging personalized medicine innovation ecosystem. His work has been published in leading interdisciplinary journals, spanning the diverse fields of medicine, biotechnology, and digital innovation. Andrew is part of a national network of innovation scholars evaluating ecosystem models that accelerate the trajectory of science and technology innovations to foster strong economic development in North America.

https://doi.org/10.1017/beq.2024.10 Published online by Cambridge University Press