# BOOK REVIEW

**Human-Centered AI**, Ben Shneiderman (2022) 400pp., £20.00 hardback, Oxford University Press, Oxford, ISBN: 9780192845290

Let us start with the book's cover. What a relief! No white, shiny, humanoid robot. No blue, floating, electronic brain. No *Sistine Chapel* handshake. Instead, the cover shows diverse people interacting with each other and with the world. Shneiderman, an expert in human–computer interaction (HCI), is sensitive to visuals and what they communicate. He briefed the cover design with care: he proposed depicting diverse people with regards to gender, age, skin color, and abilities; showing them working together and using technologies to connect and empower them; and including elements from the natural world, such as plants, birds and animals (Shneiderman, 2022).

The book's content does a good job as well: it introduces several key ideas for human-centered AI (HCAI) in an accessible manner. It consists of five parts. Part 1 provides a bit of history and context. Part 2 introduces several new views on HCAI and is, for me, the best part. Part 3 discusses four metaphors for designing and using HCAI. Part 4 discusses governance structures that would help to promote HCAI. The last part explores several potential pathways towards the future.

Overall, Shneiderman writes in an optimistic tone. In the first pages, he writes that 'a bright future awaits AI researchers, developers, business leaders, policy-makers, and others who build on AI algorithms by including HCAI strategies', and about 'supertools that amplify human capabilities, empowering people in remarkable ways' (pp.3, 4). My guess is that he chose to provide some counterbalance to authors who draw attention to the downsides and risk of the technology; one can think of Jaron Lanier, Evgeny Morozov, Safiya Noble or Shoshana Zuboff. A bit further on, he suggests finding a balance between utopia and dystopia to chart 'a path between utopian visions of happy users, thriving businesses, and smart cities, and the dystopian scenarios of frustrated users, surveillance capitalism, and political manipulations of social media' (p.12). Let us look at each of these parts in some more detail.

Part 1 discusses several ideas to provide a context for HCAI. Shneiderman writes about how HCAI draws on the traditions of rationalism and empiricism. He associates people in the AI community with Aristotle's rationalism, and people in the HCAI community with Leonardo da Vinci's empiricism. I tend to be critical whenever somebody uses a dichotomy. In this case, I am not sure whether it is particularly helpful. I would rather focus on the interactions and potential synergies between these two traditions. And I would have chosen another philosopher for rationalism; e.g., Plato. Also, in the discussion of the historical context of AI, I would have pointed at cybernetics, which coexisted with AI for a while and was then largely overshadowed by it (I will briefly return to this).

In Part 2, Shneiderman introduces a two-dimensional framework, which he also published in a 2020 paper (Shneiderman, 2020). I have found this framework very useful for discussing meaningful human control (MHC) (Steen *et al.*, 2022). MHC refers to the ambition or the requirement that people can have effective, and indeed meaningful, control over a partially autonomous system – a particularly thorny topic at play in autonomous vehicles and autonomous weapons. Shneiderman critiques a commonly used one-dimensional view on control and autonomy of self-driving cars (p.49), and instead proposes a two-dimensional understanding of control and autonomy. He draws a diagram with human control on the vertical axis, and computer automation on the horizontal axis. This creates a grid with four quadrants: the bottom left is for low human control and low computer automation; e.g., a music box which plays the music that you choose. The bottom right is for low human control and high computer automation, which we can find in an airbag, which is supposed

DOI:10.13169/prometheus.38.2.0262

to go off automatically, without human control. The top left is for high human control and low computer automation; e.g., a bicycle, which requires the person riding it to acquire and exercise skill; and the top right is for high human control and high computer automation. This last quadrant is most relevant for Shneiderman's discussion of HCAI; it refers to an optimal combination of human control and computer automation. Critically, he also discusses two regions at the right and top edges of this diagram, regions with 'excessive automation' and with 'excessive human control'. This made me think of Aristotle and his invitation to find an appropriate mean: in this case between too little and too much human control, and between too little and too much computer automation.

This two-dimensional framework enables people who are involved in the design, application and use of AI systems to discuss control and automation carefully. They can look at the framework and discuss options to go up or down (give people more or less control), or to go left or right (delegate more tasks to the system). This gives them more explicit and more nuanced options, compared with merely going up or down on a one-dimensional scale of so-called 'autonomy'.

In the top-right quadrant, Shneiderman also discusses reliability, safety and trustworthiness. He discusses these concepts in relation to design and development processes regularly used in the industry. This adds greatly to the practical applicability of the book. He discusses reliability in relation to project teams and their practices; e.g., software engineering workflows and verification and validation testing. Safety is linked to the level of the organization; e.g., to leadership's commitment to safety, and to reporting failures or near misses. And trustworthiness is discussed in the context of specific industries; e.g., ways to organize external audits, and recommendations and best practices for professional organizations.

Furthermore, Shneiderman distinguishes three categories of applications with different levels of risks: consumer and professional applications, such as recommender systems, e-commerce services, social media platforms and search engines (low-risk); consequential applications in medical, legal, environmental or financial systems that can bring substantial benefits and harms (medium-risk); and life-critical applications, such as cars, airplanes, trains, military systems, pacemakers and intensive care units (high-risk) (p.79). I would have liked to read more about these categories and particularly about the upcoming European Union's AI Act, which, for better or for worse, also works with risk categories: unacceptable risk, high risk, limited risk and minimal or no risk (European Commission, 2022).

In addition, it occurred to me that the examples Shneiderman gives are skewed towards the low-risk category: a thermostat (p.72), household appliances (p.73) and digital cameras (p.74). Elsewhere, there are many other examples of low-risk applications – the robot dog AIBO, the vacuum cleaner Roomba – and few examples of medium- or high-risk applications – a surgical system (p.109) – and societal impacts of the application of AI systems on, say, employment (pp.33–7). I see this as a missed opportunity. I would like to have seen more examples and discussions of HCAI applications with medium or high risk.

Part 3 introduces and discusses four design metaphors: intelligent agents and supertools, teammates and tele-bots, assured autonomy and control centers, and social robots and active appliances. I found these categories a bit confusing. There can be different applications within one category; e.g., an intelligent agent which is like a thinking machine, and a supertool which is meant to extend abilities (figure 11.1, on p.90). I understand agents and tools as rather different. I associate agents with machines that have some type of agency; and I associate tools with people's agency, which can be enlarged by tools. Moreover, I wonder how these metaphors may fit the HCAI framework. Can we map the metaphors into the four quadrants? Do supertools have higher human control than social robots? Or do tele-bots have less computer automation than assured autonomy? I notice that social robots and active appliances receive more attention than the other three metaphors (20 pages compared with five or six pages for the other metaphors).

I found myself speculating about interesting links that could have been made to the field of cybernetics, which emphasizes the complex and interactive relationships between people and machines and the world, and to fields adjacent to HCI, such as computer supported cooperative

work (CSCW), which is concerned with ways in which technology can support people working together. I guess that such links could help readers to understand better the ways in which AI systems are deployed in practice, embedded in complex sociotechnical systems to mediate interactions between people and between people and the world. I can imagine that views and ideas from cybernetics or CSCW can very well be used to develop and support a human-centered approach to AI.

Part 3 also offers a discussion of science goals associated with AI (p.93), and innovation goals associated with HCAI (p. 95). As with the dichotomy of rationalism and empiricism in Part 1, I wonder about the added value of such a dichotomy here. In practice, science and innovation are often combined; e.g., when people build prototypes to conduct scientific experiments, or when they use insights from scientific studies to develop and engineer products.

Part 4 discusses ways to promote reliability, safety and trustworthiness, and touches upon governance structures. Interestingly, Shneiderman relates these three concepts to different levels of abstraction: reliability to project teams and their engineering practices (e.g., retrospective analyses of failure); safety to organizations and their management strategies and cultures (e.g., leadership commitments to safety); and trustworthiness to industries and certifications and audits (e.g., by external audits). Regulation is then positioned at the level of government agencies and regulation.

Shneiderman concludes by exploring several potential pathways towards the future: boosting citizen science; stopping misinformation; and finding new treatments and vaccines. These sections contain interesting ideas. However, the discussion of citizen science avoided addressing issues of power and power differences, as if science is value neutral and unproblematic. In the context of HCAI, I would have expected a couple of remarks on the ways in which citizens can conduct science in ways that can indeed empower them. Otherwise, it is easy for industries to exploit citizens as means to collect data in ways that effectively give industries power and disempower citizens.

The book has one key shortcoming: it has little to say at any depth on ethics. For sure, there are discussions of topics that relate to ethics, such as responsibility, fairness, explainability (pp.54, 80) and bias (pp.160–4), but there is little philosophical depth. References to such ethical traditions as consequentialism, deontology, relational ethics or virtue ethics are missing. Readers only get one quote from Virginia Dignum (p.87) and a short bit about the work of Shannon Vallor (pp.259–60). Somewhat similarly, I found relatively little depth in the discussion of AI technologies. Shneiderman mentions generative adversarial networks (GANs), convolutional neural networks (CNN), recurrent neural networks (RNN), inverse reinforcement learning (IRL), but says little more about them. There is no comment on their relative strengths and limitations and only a short and general discussion on the effects of algorithms (pp.14, 160–1), drawing from O'Neil (2016).

I also missed a more thorough discussion of ways to move from ethical principles to ethical practices, though the title of chapter 18 is 'How to bridge the gap from ethics to practice'. Shneiderman's examples deal with practices or topics that are adjacent to ethics (such as project management, organizational culture and industry standards). These practices and topics can enable ethical reflection, inquiry and deliberation. They are conditions for ethical practices. But the effective promotion of ethical reflection, inquiry and deliberation requires more than having these conditions in place. I can imagine that Shneiderman, with many years of experience, would be able to give interesting examples of how to bridge this gap between principles and practices. He misses his opportunity.

Now, maybe my expectations are too high. I mean, the book does cover a broad terrain, it does a great job in promoting HCAI, putting human and societal needs center stage in the design and application of AI, and in presenting and discussing several very practical ideas – notably, the two-dimensional framework of control and automation. Moreover, the book's relatively shallow treatment of ethics is hardly unique. Many presentations of ethics in the context of technology do not go much further than admitting that privacy is important, where privacy refers to data protection, and where ethical refers to preventing bias in an algorithm's training data. Ethics can be much more than data protection and preventing bias (see Vallor, 2016: Dignum, 2019; Coeckelbergh, 2020; and, if self-promotion is allowed, Steen, 2022).

**References**

Coeckelbergh, M. (2020) *AI Ethics*, MIT Press, Cambridge MA.

Dignum, V. (2019) *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer Nature, Cham Switzerland.

European Commission (2022) 'A European approach to artificial intelligence', available at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed May 2022).

O'Neil, C. (2016) *Weapons of Math Destruction*, Penguin, Harmondsworth, UK.

Shneiderman, B. (2020) 'Human-centered artificial intelligence: reliable, safe and trustworthy', *International Journal of Human–Computer Interaction*, 36, 6, pp.495–504.

Shneiderman, B. (2022) 'Behind the cover: human-centered AI', OUPblog, available at: https://blog.oup.com/2022/03/behind-the-cover-visualizing-human-centered-ai/ (accessed May 2022).

Steen, M. (2022) *Ethics for People who Work in Tech*, Routledge, New York.

Steen, M., van Diggelen, J., Timan, T. and van der Stap, N. (2022) 'Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives', *AI and Ethics*, 18 May, available at https://link.springer.com/article/10.1007/s43681-022-00168-2#citeas (accessed May 2022).

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

*Marc Steen*
*TNO, The Hague – New Babylon, The Netherlands*
*marc.steen@tno.nl*