This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is published in Metascience, and is available online a thttps://link.springer.com/article/10.1007/s11016-023-00837-w

On Human-Centered Artificial Intelligence

Gloria Andrada, Universidade NOVA de Lisboa, gandrada@fcsh.unl.pt.

Ben Shneiderman: Human-centered AI. Oxford: Oxford University Press, 2022, 400 pp, £20 HB

Artificial Intelligence (AI) has entered our lives, and it seems that it is here to stay. However, there are different ways in which its growing integration into most aspects of our existence can happen. Ben Shneiderman's *Human-Centered AI* offers a comprehensive introduction to the human-centered framework for AI (HCAI). The main goal of the HCAI framework is to engage in AI design, innovation, and research in such a way that human needs and values take center stage. This entails designing AI systems that amplify, augment, and empower human abilities, and that promote human self-efficacy, creativity, and responsibility.

One of the driving ideas of Shneiderman's HCAI framework is that automation is compatible with human control. According to the author, their alleged incompatibility is a well entrenched and limiting belief that we all, and designers in particular, should free ourselves from. In *Human-Centered AI*, Schneiderman advocates for a two-dimensional HCAI framework, according to which achieving high levels of human control and high levels of computer automation is possible. Determining the correct balance of AI automation and human control will depend on the tasks at hand, but an important take-home message of this book is that even in conditions of high automation, a high degree of human control is possible and, in many cases, desirable. This ranges from what the author calls "lightweight" consumer applications, such as a cell phone camera, to consequential or life-critical applications, such as AI systems that are used to decide who gets a mortgage, or in self-driving cars. The key lies in prioritizing human autonomy, welfare, and safety. This means that instead of endowing technologies with high automation and low

human control simply “because we can” (p. 218), enhancing human control whenever suitable should be a priority.

To illustrate the HCAI framework and motivate its adoption by designers and other AI stakeholders, Shneiderman invites us to expand our imagination by considering different design metaphors. Instead of simply thinking of AI systems as autonomous intelligent agents, HCAI encourages us to think of them as supertools, i.e. tools aimed at extending human abilities, empowering users, and enhancing human performance. Instead of aiming at designing autonomous computers that work as “teammates”, or even anthropomorphic social robots, HCAI invites us to look for inspiration in the metaphors of tele-bots and active appliances. Finally, instead of aiming at assured AI autonomy, HCAI emphasizes human autonomy through well-designed control panels. Overall, the idea is to combine scientific and innovation goals, and take advantage of the unique features of both computers (which include sophisticated algorithms) and humans. Throughout the book Shneiderman insists on the fact that humans and computers are different, and that the former should not be a model for the latter. Ultimately, the goal should be to design computer applications that are reliable, safe, and trustworthy.

Shneiderman, who has ample experience in the field of computer science, and is a pioneer in the field of human-computer interaction, offers extensive guidelines for making AI more ethical. He motivates his theoretical background—a kind of humanist empiricism—and also offers a roadmap for bridging the gap from ideas to actual practices. HCAI requires a coordinated effort among stakeholders, including software engineers, business managers, independent oversight committees, government regulators, and also users and consumers who would benefit from taking a more active stance in their interactions with technologies. Good design should also be guided by extensive research, the retrospective analysis of failures, and continuous monitoring.

Overall*, Human-Centered AI* conveys an optimistic message and fosters “a positive mindset” (p. 225) concerning the future of AI: an optimism that is not naive, as it emphasizes many challenges that AI innovation faces (e.g. biases, intervention by malicious actors, and unethical business models).

The book will be of interest to anyone interested in AI (including software engineers, designers, computer scientists, policy-makers and philosophers) and in

our future. Its writing style is accessible, and consequently can be read both by experts and by novices. It may also be useful for pedagogical purposes. It is divided into five parts that can be read independently, as they each have an introduction that situates its content in the overall argument of the book. However, my recommendation is to follow the narrative line intended by the author.

As a philosopher interested in human nature and, therefore, in technology, I will offer three comments on Shneiderman's HCAI framework before concluding this review.

Human-Centered AI makes a case for AI systems that amplify and extend human abilities and performance. But it can be controversial what exactly we mean by that. For example, being able to do more, and being able to do things more easily and effectively should not always be considered an amplification of our abilities, or at least we should be cautious before making that claim. In many cases, instead of knowing how to do more things, we end up outsourcing our abilities to a computer, and this might entail that we are losing (instead of amplifying) our flexibility of action (Andrada 2022). That said, abilities require self-regulation: that is why I agree with the author that designing good interfaces that enhance human control is crucial for extending human agency. But we also need to critically assess the kinds of technologies that we interact with. In particular, it is important to bear in mind that our technologies and tools transform not only the nature of the tasks that we need to perform, but also the nature of our cognition itself (Menary and Gillett 2022). For example, there is growing evidence of the not-always-positive effect that GPS devices have on our navigational abilities (Gillett and Heersmink 2019). That is why, instead of assuming that the extension of our performance is always an extension of our agency, we need a more nuanced analysis. To this end, having more information available concerning the effect that specific technologies have on our cognitive capacities should be of utmost importance in an HCAI framework. This will partly include what I have elsewhere called *transformational* AI transparency (Andrada, Clowes, and Smart 2022).

My second comment challenges Shneiderman's optimistic vision of HCAI, and in particular, its role in protecting the environment. According to the author, some broader goals of HCAI are to ensure "privacy, increase cybersecurity, support social justice, and protect the environment" (p. 56). I would like to share his optimism, but I

wonder to what extent AI (whether human-centered or not) might also be part of the problem, and not just the solution. Shneiderman does acknowledge some of the risks that AI entails, but he does not consider the ecological footprint of AI systems, or the importance of gathering more information on that matter (see Strubell, Ganesh, and McCallum 2019).

Moreover, we find almost no reference to AI hardware or the actual material implementation of AI systems. It is true that when considering the voices of skeptics, Shneiderman does refer once to the extractive nature of AI, "which threatens the environment, fosters unhealthy working conditions, and mistreats users" (p. 225). But his optimistic spirit quickly soothes us with the affirmation that "more positive outcomes are possible" (p. 225). I also think that better outcomes are possible, but to achieve them, or to get somewhat closer to a better world, we cannot forget the materiality of AI systems. This is a key dimension of AI transparency that I have elsewhere called "material transparency" (Andrada, Clowes, and Smart 2022), and I believe it should be addressed in a HCAI framework.

This brings us to my third and final comment, which focuses on the "human" in human-centered artificial intelligence. We should not forget that talking about a universal human is problematic, as we might easily occlude differences and impose a way of understanding humanity that not every social group or culture shares. To be fair, Shneiderman does mention in several places the importance of being aware of cultural differences, but he does seem to assume that human self-efficacy and autonomy are universal values. I wonder to what extent that is correct, and even if it is, there are different ways in which those values can be understood and actualized. For example, under a certain system of beliefs, human self-efficacy and autonomy might be the ultimate goal, and that might entail certain forms of technological innovation. However, from a different perspective, sustainability and care for diverse forms of life might be more or as important. From this perspective, achieving human autonomy might entail doing things radically differently. Overall, non-dominant human cultures might have a different way of understanding humanity and what is valuable, and these might imply different technological designs and innovation. Consequently, deep reflections are still needed on what human-centered AI entails and requires.

References

Andrada, Gloria. 2022. Extending knowledge-how. *Philosophical Explorations*. doi: 10.1080/13869795.2022.2116090

Andrada, Gloria, Robert William Clowes, and Paul Smart. 2022. Varieties of transparency: Exploring agency within AI systems. *AI & Society*. doi: 10.1007/s00146-021-01326-6

Gillett, Alexander, and Richard Heersmink. 2019. How navigation systems transform epistemic virtues: Knowledge, issues and solutions. *Cognitive Systems Research* 56: 36-49.

Menary, Richard, and Alexander Gillett. 2022. The tools of enculturation. *Topics in Cognitive Science* 14 (2): 363-387.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. arXiv archive. https://arxiv.org/abs/1906.02243. Accessed [December 10, 2022].