Supporting human flourishing by ensuring human involvement in AI-infused systems

Joel Chan, Hal Daumé III, John P. Dickerson, Hernisa Kacorri, and Ben Shneiderman
{joelchan, hal3, johnd, hernisa, bshneide}@umd.edu
Human-Computer Interaction Lab
University of Maryland, College Park

Abstract

Researchers, developers, business leaders, policy makers and others are expanding the technology-centered scope of Artificial Intelligence (AI) to include Human-Centered AI (HCAI) ways of thinking. This expansion from an algorithm-focused view to embrace a human-centered perspective, can shape the future of technology so as to better serve human needs. Educators, designers, software engineers, product managers, evaluators, and government agency staffers can build on AI-infused technologies to design products and services that make life better for the users. By switching the scope from technology-centered to human-centered, we can build AI-infused tools that enable people to better care for each other, build sustainable communities, and restore the environment.

1 Introduction

A new synthesis of disciplines is emerging, in which AI-based intelligent algorithms are combined with human-centered design thinking to make Human-Centered AI (HCAI). This synthesis of disciplines increases the chance that technology will empower rather than replace people. In the past, researchers and developers focused on building AI algorithms and systems, stressing machine autonomy and measuring algorithm performance. The new synthesis values AI and gives equal attention to human users and other stakeholders by raising the prominence of user experience design and by measuring human performance. Researchers and developers for HCAI systems value meaningful human control, putting people first by serving human values such as rights, justice, and dignity, thus supporting goals such as self-efficacy, creativity, responsibility, and social connections [20].

The higher level goal is to support the 17 United Nations Sustainable Development Goals (`https://sdgs.un.org/goals`), which were established in 2015 to set aspirations for 2030. These goals include elimination of poverty, zero hunger, quality education, and reduced inequalities. Other ambitions address environmental issues such as climate action, life on land, life below water, and sustainable cities and communities.

These ambitious goals will be realized primarily through human behavioral changes, which can be aided—but hardly replaced—by technology: they all have behavioral components, demand attitude changes, and require commitments to energy and money. For this reason, we argue that a fundamental shift is needed in how we design AI technologies: we need to move away from the vision of "autonomous AI systems" towards *AI-infused tools* that are not only human-centered in the sense of aligning with human values, but also human-centered in the specific sense of centering human capabilities and involvement.

In this position paper, we articulate a four-fold strategy to designing AI technologies that center human capabilities and involvement: 1) changing design metaphors around AI technology development, 2) developing user interface design guidelines for foregrounding human-control of AI-infused tools and

Human Centered AI Workshop, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

active appliances, 3) developing holistic management via AI operations and control centers, and 4) developing human-centered objective functions and datasets for AI models. These four strategies are not an exhaustive listing; instead, we offer them as suggestions for research directions we have found to be compelling in our shared and respective research trajectories with AI-infused systems across diverse areas such as visualization and knowledge discovery, markets and optimization, accessibility and education, natural language processing, and interactive creativity support systems. Our goal is to start a fruitful dialogue around how to best achieve the ambitious goal of aligning these systems with sustainable development goals.

2 Changing Design Metaphors

The first key strategy we propose is to adopt new Design Metaphors that can be combined with the existing ones [18]:

1. from intelligent or autonomous agents (a common focus at NeurIPS) to *AI-infused tools*
2. from teammates to *tele-bots*
3. from autonomous systems to *control centers*
4. from social robots to *active appliances*

The initial set of metaphors — as intelligent or autonomous agents, teammates, autonomous systems, and social robots — are enticing: journalists, headline writers, graphic designers, and Hollywood producers are entranced by the possibilities of robots and AI. But the second set of metaphors are necessary for shaping attitudes and expectations towards a view of AI as tools in support of human control and responsibility. For example, changing the design metaphor from intelligent agent to AI-infused tools can appropriately shift design goals away from ideas of thinking machines as cognitive actors, towards goals of extending abilities, empowering users, and enhancing human performance.

With this fresh thinking, researchers, developers, business leaders, and policy makers can find combined designs that will accelerate HCAI thinking. A greater emphasis on HCAI will reduce unfounded fears of AI's existential threats and raise people's belief that they will be able to use technology for their daily needs and creative explorations. It will increase benefits for users and society in business, education, healthcare, environmental preservation, and community safety [19].

3 User Interface Guidelines for Ensuring Human Control While Increasing Automation

A second key strategy is to design user interfaces and control panels that ensure meaningful human control while increasing the level of automation. These interfaces and control systems should give users of AI-infused tools and active appliances greater understanding of the state of the machine and what it will do next or how it can fail, as its behavior can be alien [22]. Users are in control when they have access (*e.g.*, via inclusive visual, auditory, or haptic previews) to what their computer could do, so they can select from alternatives and initiate actions, then follow it through during execution. This is what users of digital cameras and navigation systems already have, but the guidelines need to be applied in other applications. Such design principles are particularly important when considering users whose expertise varies; some may desire simpler, "customer-friendly" interfaces, while others can be wary of anything that appeared automated [2]. Similar designs for industrial robots, drones, financial trading systems, ship navigation, and medical devices follow the Human-Control Mantra: Preview first, select and initiate, then view execution.

Another aspect of this approach will involve innovations that lower the barriers for users to directly influence the AI models that are supporting their tools. Such innovations might include expansions of the *machine teaching* paradigm, where the focus is given to "making the process of teaching machines easy, fast and above all, universally accessible" [21]. This paradigm can have further implications for inclusion and personalization [13]. For instance, teachable interfaces can enable blind users to train an object recognizer with their photos [14]. Participatory in nature, by exerting control over the behavior of AI models that underpin AI-infused tools, machine teaching allows for observations and reflections and promotes user experimentation that can spark counterfactual thinking [6] for adults [12] and children [9].

4 Holistic Management via AI Operations & Control Centers

A third key strategy, dovetailing with that above, is to enable operations and control *centers* that empower teams of humans to comprehend, change, and control the behavior of sets of AI-infused tools. For decades, industrial and government management of critical infrastructure has resulted in both physical and virtual *operations centers* wherein humans continuously monitor a system for performance. Examples include: Network Operations Centers (NOCs), where information technology (IT) teams monitor the health of a company's network; Tactical Operations Centers (TOCs), where military personnel monitor active and passive tactical elements of a mission; and, Security Operations Centers (SOCs), where teams monitor organization and technical security for a firm. In this same vein, operations centers for AI/ML have begun to appear to support the burgeoning ML Operations (MLOps) role, and serve as a centralized setting to deploy, monitor, and maintain production machine learning models. Yet, these centers typically focus on monitoring simple metrics such as accuracy or concept drift, providing alerting services based on simple learned or human-set threshold-based rules, and provide a short-term focus on keeping machine learning models running smoothly. Extensions to the MLOps paradigm, such as ModelOps [3], wrap lightweight concerns of the firm around model management, but are still fundamentally performance-metrics-driven and thus focus more on short-term operations and less on human control.

Rigid measures of accuracy serve a purpose, reflecting one view of a model's current performance; yet, as discussed in the succeeding section, it is important to consider a broader scope of human-centered metrics and objectives that reflect the needs of downstream human stakeholders. Indeed, in line with recent guidance from the US National Institute of Science and Technology (NIST) and other calls for AI Governance [17], it is necessary to widen the scope of present-day AI/ML Operations Centers to include broader classes of metrics and objectives.

5 Human-centered Objective Functions and Datasets

A fourth key change is to revisit the objective functions that are used to train AI models, to move them away from rigid accuracy measures and emphasis on autonomous action, to be driven much more fundamentally by human-centered criteria. These objective functions can be significant modifications to familiar functions, such as thoughtfully incorporating considerations of fairness, diversity, or equity in the objective functions of matching algorithms that undergird marketplaces. Indeed, some progress has been made melding preference and value judgment aggregation techniques from the computational social choice [5] literature into AI applications [e.g., 16, 10], yet those learned objectives are still brittle and may not accurately capture stakeholders' wants. Some may be adaptations of existing objectives, accounting, for instance, for the harms incurred by language processing systems built with an implicit or explicit a binary lens on human gender [7]. Others might be entirely new, such as optimizing for certain kinds of "noise" or surprise in output, which may be especially useful for AI-infused tools that augment human creative capacities, such as systems for synthesizing contributions to large-scale collaborative innovation platforms to inspire further exploration of a solution space [8].

In addition to the development of human-centered metrics, it is equally important to pair these with datasets constructed with human-centered values in their design. In order to contribute meaningfully to the research dialogue, the human-centered values on which datasets (and metrics) are built deserve to be made explicit, carefully documented [11], and grounded in the relationship between a system and the social hierarchies in which it participates [4]. Even with good intent, datasets sourced from underrepresented communities can be a double-edged sword, calling for better sharing practices as well as technical, legal, and institutional privacy frameworks that are more attuned to concerns from these communities. For instance, data sourced from people with disabilities are crucial for inclusion and innovation but they pose privacy and ethical concerns as people with distinct data patterns may be more susceptible to data abuse and misuse, *e.g.*risks of inaccurate or non-consenting disclosure of a disability [15]. This work will likely be difficult and costly, at least at first, but the payoff to doing so will be well worth it: a closer alignment of fundamental AI performance with human values.

6 Conclusion

These ideas for human-centered AI are still nascent and will need to be refined in practice, tuned to the needs of each industry, and adjusted as innovations emerge. They are gaining acceptance, but there is

still resistance from those who believe in established ways of working. This vision for the future is still a minority position, so there is much work to be done to steer researchers, developers, managers, and policy makers to this new synthesis. Other challenges come from the numerous threats such as misinformation, cyber-crime, political oppression, and online bias, hatred and bullying—oftentimes made worse by deployed AI, motivating a parallel thread of thought understanding the multifaceted role computing writ large can play in society [1].

However, we feel confident that the future is human-centered – filled with AI-infused tools and active appliances that amplify, augment, and enhance human abilities, empowering people in remarkable ways while ensuring human control. This compelling HCAI prospect enables people to see, think, create, and act in extraordinary ways, by combining engaging user experiences with embedded AI algorithms to support services that users want. The HCAI prospect contributes to hope-filled agendas for healthcare, community safety, economic development, racial justice, accessibility, and environmental sustainability.

References

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 252–260, 2020.

[2] David Alvarez-Melis, Harmanpreet Kaur, Hal Daumé, III, Hanna Wallach, and Jennifer Wortman Vaughan. From human explanation to model interpretability: A framework based on weight of evidence. In *HCOMP*, 2021.

[3] Soyeb Barot. A guidance framework for operationalizing machine learning, 2018. Gartner Research Report.

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé, III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2020.

[5] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[6] Ruth Byrne. *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press, Cambridge, MA, 2007.

[7] Yang Trista Cao and Hal Daumé, III. Toward gender-inclusive coreference resolution. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2020.

[8] Joel Chan, Steven Dang, and Steven Dow. Comparing different sensemaking approaches for large-scale ideation. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[9] Utkarsh Dwivedi, Jaina Gandhi, Raj Parikh, Merijke Coenraad, Elizabeth Bonsignore, and Hernisa Kacorri. Exploring machine teaching with children. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 2021.

[10] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. In *Communications of the ACM*, 2021 (to appear).

[12] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Hernisa Kacorri. Teachable machines for accessibility. *SIGACCESS Access. Comput.*, (119):10–18, November 2017.

[14] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 5839–5849, New York, NY, USA, 2017. Association for Computing Machinery.

[15] Rie Kamikubo, Utkarsh Dwivedi, and Hernisa Kacorri. Sharing practices for datasets related to accessibility and aging. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

[17] Reva Schwartz, Leann Down, Adam Jonas, and Elham Tabassi. Draft NIST Special Publication 1270: A proposal for identifying and managing bias in Artificial Intelligence, 2021. Draft publication. `https://doi.org/10.6028/NIST.SP.1270-draft`.

[18] B. Shneiderman. Design lessons from ai's two grand goals: Human emulation and useful applications. *IEEE Transactions on Technology and Society*, 1:73–82, 2020.

[19] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36:495 – 504, 2020.

[20] Ben Shneiderman. *Human-Centered AI*. Oxford University Press, 2022, forthcoming.

[21] Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. Machine teaching: A new paradigm for building machine learning systems, 2017.

[22] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, May 2019.