npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01415-y

A caution against customized AI in healthcare

Kristin M. Kostick-Quenet

This article critiques the shift towards personalized AI in healthcare and other high-stakes domains, cautioning that without careful deliberation, customized AI systems can compromise the diversity and reach of human knowledge by restricting exposure to critical information that may conflict with users' preferences and biases. Customized AI should be applied with caution and intention where access to a wide and diverse range of perspectives is essential for impartial, informed decision making.

The trend of personalizing artificial intelligence (AI) systems is gaining momentum as AI companies compete to develop tools with broad consumer appeal and respond to challenging ethical questions about how best to prioritize and align content with diverse human values and beliefs. AI-powered chatbots and answer engines are quickly replacing more traditional search engines as a preferred platform for acquiring knowledge, placing growing pressure on AI developers to present outputs that resonate with what consumers believe to be accurate representation of the "truth." However, the wide diversity of belief systems and perspectives among users poses significant challenges in deciding which information to prioritize, especially where truths are contested or where evidence diverges. This is especially relevant in healthcare, where stakes are high for pinpointing the precise nature and severity of pathology and effective treatment pathways. A primary utility of AI in healthcare is to identify patterns that are difficult for clinicians to see on their own (e.g., due to human imperceptibility, insufficient resolution, voluminous data); however, they rarely point to a single truth and more often present a range of probabilities. As such, AI systems designed to augment clinical decision making face the same problems of acceptability and uptake as consumer-grade AI-powered chatbots and answer engines. AI developers in healthcare (as in other fields) are similarly responding to these acceptability challenges by offering systems that customize outputs to users' (e.g., physicians') style and content preferences. What follows is a critique of this shift and a call for applying customized AI with caution and explicit intentionality in critical domains like healthcare where access to a wide and diverse range of perspectives is necessary for impartial, informed and responsible decision making.

Background: A push for personalized AI

In 2022, AI moved into the global spotlight after the release of OpenAI's ChatGPT, a large language model (LLM) that generates responses according to a complex calculus of word sequence, context, and coherence. Importantly, these calculations depend strictly on the content, quality, and quantity of training data sets and have no inherent relationship to any objective truth. Given that LLMs are trained on vast portions of the internet, critics[1] argue that AI outputs mirror biases embedded in online texts leading to overrepresentation of dominant (e.g., English, Western) viewpoints. The AI community has responded by refining datasets and employing techniques like fairness constraints and adversarial training to ensure more balanced representation of perspectives. However, recent controversies (e.g., over Google Deepmind's AI assistant *Gemini* depicting America's "founding fathers" as people of color) reveal that some developers are overshooting these attempts, prioritizing fairness and non-bias at the expense of accuracy. Such controversies highlight the need to effectively address algorithmic bias, but also highlight a deeper question about how to shape AI systems that, in turn, shape the content of information we consume. AI systems are rapidly becoming essential tools for human reflection, influencing our understandings, opinions, and decisions around entertainment, leisure, work, and even critical domains like healthcare.

Despite their far-reaching political and social implications, important AI design choices are being left to Big Tech companies that are more directly incentivized to expand their user base than to expand human knowledge. Big Tech's response to the consumer value alignment problem is captured by what Open AI's Sam Altman has described as a "simple" solution: Let users decide. Rather than navigate a complex landscape of contested worldviews, AI developers are pioneering a new generation of AI systems that learn from users' feedback to deliver content that aligns with a user's specific style and content preferences. This design choice absolves developers of the task of deciding which information to prioritize, while users receive content they perceive to be more relevant interesting, thereby enhancing the product's appeal.

The rapid adoption of algorithmic customization (or "personalization," used here interchangeably) is demonstrated by the latest trend of enhancing systems' capacity to remember and adapt. Expanding the number of tokens per context window allows AI systems to better recall what users said at the beginning of a conversation or even previous conversations. Researchers[2] are also working on "infini-attention" models that leverage data compression to enable infinite memory. The vision is to offer users a bespoke, lifelong companion, powered by their entire digital footprint, promising to "never forget what's important to you" and always be "on your side"[3]. Mustafa Suleyman, co-creator of Inflection AI's new

Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX, USA. e-mail: kristin.kostick@bcm.edu

"personalized intelligence" system, forecasts a future with millions of personalized AI systems, serving not only individuals but also businesses and even governments.

While AI customization helps to address the interrelated challenges of value alignment, acceptability, and uptake, lessons from two decades of social media experimentation should caution us about letting algorithms shape the range and diversity of information consumed by humans. As we enter the era of "personalized everything," AI developers and policymakers seem to overlook the social and epistemic implications of granting algorithms (beyond those used in social media) the power to curate human understandings of the world, and the risks of indiscriminately applying customization across critical sectors like healthcare, banking, education, transportation, manufacturing, agriculture and energy, where customized AI is already appearing. While algorithmic filtering and curation may seem innocuous, greater attention from researchers and policymakers is needed to assess the risks of allowing AI systems to increasingly mediate human understandings and perspectives.

Evolution of personalized AI

The evolution from one-size-fits-all to personalized AI models represents a significant shift in the objectives and capabilities of AI technologies. Early AI systems were confined to fixed, narrow, and pre-defined objectives. However, the landscape of AI has radically changed with advancements in inverse reinforcement learning (IRL) that enable AI systems to operate with a degree of uncertainty about their objectives. Unlike traditional reinforcement learning, which relies on pre-specified rewards and punishments to teach behavior, IRL allows machines to infer and learn underlying preferences and principles from observed human behavior. This capacity has led to an explosion of "recommender systems" taken up by social media and online marketers to deduce user preferences from online activity, fueling a multibillion-dollar industry dedicated to enhancing user engagement on digital platforms. Supercharged with deep learning for further accuracy and effectiveness, IRL can enhance user experience by automatically tailoring content to reduce exposure to contradictory or disquieting information, minimizing discomfort or "friction".

Regulatory landscape

While effective, these algorithms are also widely blamed for fostering "echo chambers" and "filter bubbles" that reinforce users' existing beliefs. Legislation following the Cambridge Analytica scandal, which spotlighted the risks of filter bubbles for ideological polarization and behavioral manipulation, focused on strengthening data privacy protections but did not directly address risks of narrowing or distorting information on a wide scale. Other regulations attempting to address filter bubbles directly (the Filter Bubble Transparency Act of 2021) never passed. More recent initiatives in the U.S[4]. and E.U[5]. around AI focus on combatting new forms of misinformation like AI-generated deepfakes but continue to overlook subtler risks of information filtering that can critically undermine AI's transformative potential in sectors vital to society, where friction is essential for promoting reflective thinking and awareness of diverse viewpoints and biases.

Recommendations for Policy & Development

Algorithmic constraints on accessible information can lead not only to echo chambers but also to discrimination, bias, loss of decision autonomy, dependence on AI systems, and diminished human judgement. Further, as AI systems evolve into more independent "agents" capable of handling complex tasks with minimal oversight, the impacts of customization will increasingly manifest in the real world. Rather than confining these concerns to the realms of social media or politics, AI policymakers should consider a broader range of domains where AI customization increasingly shapes human exposure to information. Efforts to make AI inference more inclusive focus on mitigating algorithmic bias or balancing representation in training data sets; but more research and policy should focus on AI model and interface designs that encourage wide-ranging and analytical perspectives, especially for systems that significantly affect human safety and well-being. Below are three recommendations for AI development and policy initiatives to help cultivate "constructive friction" in AI systems and encourage the creation of AI that not only captivates but also fosters critical thinking and broad perspectives.

Leverage decision science to optimize for critical reflection

Generative AI systems now allow users to ask questions in natural language and typically receive a "best fit" result in a seamless, conversational thread. As developers work to ensure responses are accurate, the demand for accuracy among consumers remains complex. While users may value accuracy, their online behavior suggests a stronger preference for information that aligns with their beliefs and preferences. Even this is not straightforward, as consumption habits are influenced by interface design features in ways that are not well understood[6]. The new proliferation of AI chatbots offers a natural testbed to experiment with these tendencies on a massive scale to identify forms of information delivery that promote critical evaluation of misinformation and consideration of alternative perspectives, without fundamentally compromising sustained user engagement.

A wealth of decision science and behavioral economics literature can help to design such experiments. For example, empirical research shows that the way choices are presented (i.e., "choice architecture"[7]) can significantly influence decision making. Presenting multiple distinct options aids informed decision-making, but an excess can increase cognitive load[8]. Further, humans imperfectly evaluate information and make decisions influenced by cognitive and emotional biases and contextual factors (i.e., "bounded rationality")[9]. With generative AI systems rapidly replacing traditional search engines for building knowledge, policymakers and Big Tech should partner to promote consumption of diverse content, combat confirmation bias, and foster social unity. Strategies could include integrating alternative perspectives into outputs or using design features like accordion menus, pop-up windows, and progressive disclosure buttons to enhance awareness of contextual information. A vast array of experimental designs from user experience (UX) testing (e.g., A/B testing; conversion rate optimization; multivariate testing) can help to identify decision architectures with positive versus negative impacts on critical thinking.

Ensure intentionality

Not all customization is harmful. It is essential to differentiate between using algorithmic customization approaches (e.g., unsupervised learning, latent pattern mining, anomaly detection, dimensionality reduction, etc.) to *discover* verifiable knowledge versus to adapt the *delivery* of knowledge to ensure palatability and user acceptance. This distinction (illustrated in Fig. 1) can help evaluate the rationale behind customization and align it with application goals.

Consider an AI-powered radiology tool that identifies relevant areas of a medical image and tailors outputs in line with a radiologist's preferences and past interactions with the system. The design choice to emphasize physician engagement is based on an appreciation that the uptake and utility of an AI tool depends not only on its performance but also on the degree to which users trust it. Emerging research suggests that user trust is partly influenced by whether an AI system considers a user's preferred variables and ways of receiving or visualizing outputs[10,11]. Leveraging these insights, novel AI systems for medical imaging analysis (e.g., Aidoc; PathAI) use customization to adapt workflows to mirror how individual radiologists prioritize cases, highlight abnormalities, and generate reports. Similarly, medical "scribes" that automate physician notetaking (e.g., DeepScribe) can be personalized to adapt to physicians' speech patterns, medical terminology preferences, and summarization styles that match a physician's preferred level of detail and prioritized information, and perform selective information retrieval (e.g., from electronic health records) based on a physician's regularly referenced information. Clinician-facing medical chatbots (e.g., OpenEvidence) designed to deliver clinically relevant literature, evidence-based guidelines, and decision support summaries at the point of care are using customization to align outputs with patterns in a

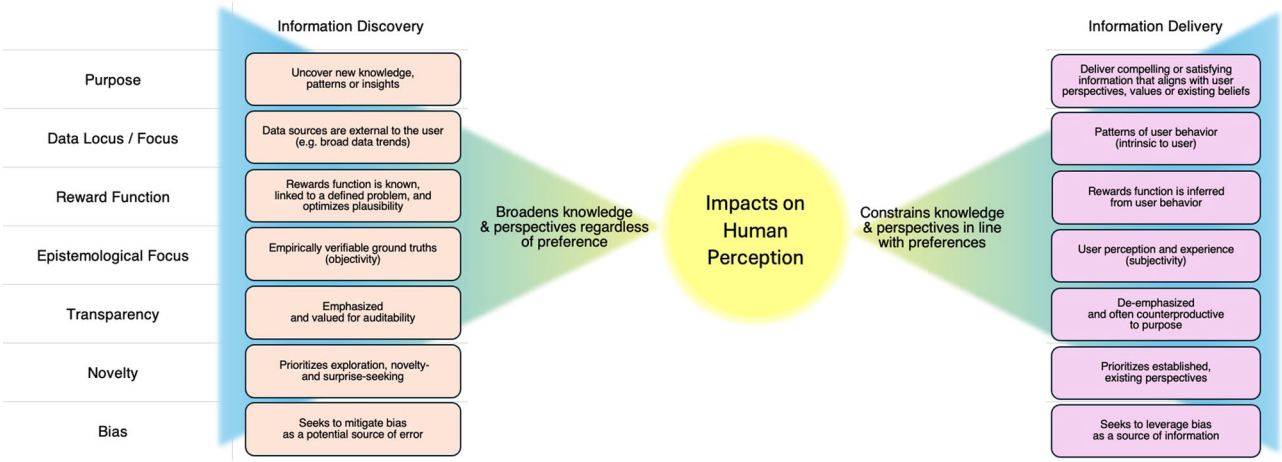

Fig. 1 | Designing AI for Information Discovery versus Information Delivery. This figure presents distinctions between AI systems customized for information *discovery* and those for information *delivery* across several dimensions and highlights the impacts of these AI design approaches on human perception.

clinician's queries and preferred sources. Other emerging AI tools in healthcare (e.g., for algorithmic risk prediction; patient scheduling; patient monitoring early alert systems) likewise feature customized information delivery.

While tailoring outputs may lead to greater uptake of tools that ultimately enhance diagnostic efficiency and precision, it may cause physicians to miss critical information if preferences do not reflect variables with the highest predictive value for treating disease, or if preferred outputs constrain a physician's ability to consider the wide range of information necessary to make informed decisions. Further, outputs from AI-based medical scribes may embody a physician's unique expertise but also their biases. These biases may then be documented and perpetuated in a patient's medical record in ways that shape (and potentially constrain) the perspectives and understandings of future physicians.

We should be designing AI tools for healthcare and other high stakes fields that work to offset, not perpetuate, human biases. Given the variability in how physicians interpret and respond to clinical information, hospital governance—as well as developers and policymakers—should establish limits on AI tool's accommodation of user preferences, balancing trust, engagement, and other objectives like accuracy or comprehensiveness. Further, hospitals should require regular audits of AI performance impacts on clinical decision making and health outcomes to ensure critical variables with high predictive value are not overlooked in favor of user preferences. These checks and balances are likewise critical in other sectors where fairness, standardization, or impartiality are principal considerations (e.g., national security; legal rulings; standardized testing). In these contexts, developers should prioritize design elements that enable users to access "ground truths"—accurate, verifiable information that reflects reality (e.g., the severity of a patient's disease, the full range of security threats). Depending on its aim, customization can either lead us closer to or further from these ground truths, wielding a power to constrain human perception as much as it can expand it.

Customization thus requires caution and intentionality, supported by rationales that align with responsible uses of AI across different domains and settings. While all forms of information delivery, even books, guide our attention, the distinguishing aspect of AI that demands closer scrutiny is the degree to which reliance on AI over other sources of knowledge will influence human perspectives over longer timescales. The widespread integration of generative AI marks the first time that humans are relinquishing responsibility and control over generating and curating knowledge to a non-human entity, with uncertain impacts on human progress. Intentionally programming these systems to restrict the information we ingest must be done with utmost care. While we have ample evidence of the harmful impacts of content filtering in social media, we do not yet know the harms of taking this approach with generative AI. Now is the time to anticipate such consequences. Developers could be incentivized to disclose the potential consequences of customization. Modeled after "social impact statements" that prompt developers to consider the effects of AI systems on different groups or individuals, "design impact statements" could encourage developers to transparently evaluate how curating information for engagement, acceptability or other reasons might inadvertently exclude important information or perpetuate existing biases. Developers could also highlight supplementary design choices to ensure that users are made aware of domain-critical information, even if it does not align with user preferences or past interactions. Clear documentation of personalization parameters across use cases can also help to build an evidence base of impacts.

Prioritize unconventional thinking

Generative AI stands out from previous iterations of AI for its significant potential to expand the boundaries of human cognition and transcend conventional modes of thought. Despite being trained on existing knowledge, GenAI can produce entirely novel content. It does this by using two main deep learning techniques, generative adversarial networks (GANs) and transformers, which together enable AI to create unique material that is not just a copy or remix of existing data. Various forms of reinforcement learning can encourage GenAI to "think outside of the box." Building "intrinsic rewards" into a system, like rewarding novelty- and surprise-seeking, can encourage GANs to explore new possibilities and deviate from existing expectations. Similarly, shaping rewards can help to gradually shift an algorithm's focus towards more complex or unexpected outcomes. Incorporating other motivation-like mechanisms indirectly through novel loss functions, regularization methods, or (e.g., curiosity-driven) training strategies can encourage diversity and exploration. These approaches are what afford these systems their unprecedented potential to expand the frontiers of human knowledge and creativity.

However, the drive to outwardly expand human knowledge lies in direct contradiction to the impulse that underlies personalization, which is to narrowly channel information in line with existing inclinations. AI policies should actively promote the capacity of AI systems to help humans critically reflect, evaluate and act on information in *new* ways, not according to the old ways. Design choices should encourage us to engage in unconventional or divergent thinking, consider alternative perspectives, and pursue explainability to better understand causality. For instance, new research[12] suggests that introducing certain forms of "prompt engineering" (crafting user query inputs to guide model outputs) can both enhance AI effectiveness and bolster users' capacity to assess and gauge system performance. These include prompt strategies that steer AI systems to decompose complex queries into stepwise instructions (instruction

generation) and to deliver outputs with stepwise explanations (chain of thought generation). Offering these intermediate reasoning steps on both sides (input/output) of a user's query can enhance transparency and explainability and enable critical evaluation of outputs.

Such an approach aligns with recent policy suggestions that AI systems provide metadata about system inputs and outputs that permit users to rigorously assess a system's performance accuracy, reliability, and fairness. Scholars have proposed solutions like putting "nutrition facts labels" on AI outputs[13], or implementing design approaches that decelerate heuristic thinking or mitigate cognitive and emotional biases that inhibit users from dispassionate evaluation of facts (e.g., racial bias) or an over-reliance on system results (e.g., automation bias)[14]. Interactive design features mentioned earlier, including expandable elements, dynamic filtering or resorting, or translation of information to other multimodal formats (e.g., text-to-image; image-to-sound) may serve mutual goals of exposing alternative perspectives while igniting human creativity and imagination.

A path forward

AI policy and development should look beyond the most nefarious forms of AI-generated misinformation to address more nuanced forms of knowledge curation that limit rather than expand human outlooks across the wide variety of domains where AI is present, from entertainment and politics to medicine, engineering, and space exploration. Vast societal resources are being dedicated to creating AI systems that augment human intellect, ingenuity and creativity. This may be better accomplished through incremental advancement towards objective ground truths, rather than uncritical celebration of subjectivity, relativism and satisfaction of personal preferences. Both trajectories can exist for AI. This paper offers actionable suggestions to balance preference-sensitive with preference-agnostic designs where truth matters most.

Received: 6 June 2024; Accepted: 22 December 2024;
Published online: 07 January 2025

References

1. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. "On the dangers of stochastic parrots: Can language models be too big?." In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623, Virtual Event, Canada ACM. https://doi.org/10.1145/3442188.3445922 (2021).
2. Munkhdalai, T., Faruqui, M. & Gopal, S. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. arXiv:2404.07143 (Preprint) (2024).
3. Replika. http://replika.com/ (2024).
4. Biden, J. R. Jr. Executive Order 14110.: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Signed 10/30/2023. Federal Register on November 1, 2023 under document number 2023-24283, vol. 88, p. 75191.
5. European Parliament and Council. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act). Official Journal of the European Union, L 168, 12 July 2024, pp. 1–60.
6. Shavit, Y. et al. "Practices for governing agentic AI systems." Research Paper, OpenAI (2023).
7. Kitchens, B., Johnson, S. L. & Gray, P. Understanding Echo chambers and filter bubbles: the impact of social media on diversification and partisan shifts in news consumption. *MIS Q.* **44**, 1619–1649 (2020).
8. Sweller, J. Cognitive load theory and educational technology. *Educ. Technol. Res. Dev.* **68**, 1–16 (2020).
9. Gigerenzer, G. "What is bounded rationality?" in *Routledge Handbook of Bounded Rationality* (Taylor & Francis Group, 2020).
10. Schrills, T. & Franke, T. "Color for Characters - Effects of visual explanations of AI on trust and observability" in *Artificial Intelligence in HCI* (Springer, 2020).
11. Aidoc. (2022). Next Gen Radiology AI: The Journey from an Algorithm to a Clinical Solution. Retrieved from https://www.aidoc.com/wp-content/uploads/Aidoc-Next-Gen-Radiology-AI-Whitepaper.pdf.
12. Nori, H. et al. Can generalist foundation models outcompete special-purpose tuning? CaḌse Study in Medicine. arXiv preprint arXiv:2311.16452 [Preprint] (2023).
13. Gerke, S. 'Nutrition Facts Labels' for Artificial Intelligence/Machine Learning-Based Medical Devices-The Urgent Need for Labeling Standards. *George Wash. Law Rev.* **91**, 79 (2023).
14. Quenet, K. K. & Gerke, S. AI in the hands of imperfect users. *NPJ Digit. Med.* **5**, 197 (2022).

Author contributions

K.M.K-Q. is the sole contributor to the conceptualization and writing of this paper.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Kristin M. Kostick-Quenet.