Human Centered AI Alone
Will Not Ensure A Path To Human Empowerment

Christine Custis[1]

[1]Institute for Advanced Study
ccustis@ias.edu

Abstract

This paper explores how we currently understand and navigate the challenges AI poses to human agency and autonomy and how our aim of governance requires adjusting as it is still not properly centering humans. We require a focus on an outcome of human empowerment or the best possible future for humans in addition to our current focus on model, machine, and algorithm safety acceptance or the least harmful future for humans. This AI Governance Workshop paper argues that authentic human-centeredness in the research, design, development, and deployment (RDDD) of AI can compliment human-compatibility and/or human-values-alignment constructs and can add the friction necessary to 1) foster human empowerment, 2) reverse heteronomy, and 3) avoid a future where the achievement of a fully embodied AI system simultaneously influences the disembodiment of humans. The paper proposes a structure and diagnostic framework for defining human empowerment.

Introduction

In May of 2025, Wired published an article announcing the groundbreaking work of a British startup whose software application, Silence Speaks, is said to have been created by and for the deaf community. The product translates text, voice, and video into British Sign Language (BSL) and offers the hard of hearing access to an AI-generated translator created as a photorealistic synthetic or animated avatar via a smartphone. The application's "model was trained with datasets covering regional dialects, contextual language, and emotional tone" (Wired, 2025) and the original full-stack engineers hired by the CEO were all sign language users. Although translation is done through the cloud and instant, on-device use is not currently offered, future versions of the product are meant to support accessibility and inclusion for the deaf community and additional services such as captioning for users who have difficulty interpreting sign language.

The overarching goal of the application leans towards the intention of human inclusion, specifically the inclusion of the deaf and hard of hearing communities and their capacity to communicate and participate in society. This is an example of an AI application intended for good and operating as expected with meaningful foundations in fair model design. But is that sufficient? Shouldn't we push past the questions of beneficence and harm reduction and go further in our quest for human empowerment? As a user explores the Silence Speaks application and deepens their reliance on its translation services, is their choice influenced or their preference constrained based on the model training or platform architecture? Will users who are not familiar with the design and development of the system be able to distinguish instances in which the tool is used to complete a communication task versus when the AI transforms their belief or understanding of a concept? What insights into and opportunities for input do human users have into the goal formulation, reasoning, and strategic adaptation of this machine learning tool? Which of our current responsible AI approaches to model testing and assessment can measure the loss of human dominance over the system? Questions like these are relevant ones for ML/AI researchers because they go well beyond a focus on model, machine, and algorithm safety acceptance or the least harmful future for humans. They help us to reframe our thinking about AI systems such that we prioritize the needs of an empowered humanity instead of simply limiting the harms to which humans are exposed.

This AI Governance Workshop paper posits that the aim of our current efforts to ensure safety in AI models does not adequately empower humans. To adjust our aim and properly center humans in the research, design, development, and deployment (RDDD) of AI, we require a focus on an outcome of human empowerment (or the best possible future for humans) in addition to our current focus on model, machine, and algorithm safety acceptance (or the least harmful future for humans).

As a concrete example of how a new paradigm for human empowerment can benefit AI, we return to the Silence Speaks software and consider actions we can take to empower humans. In this new paradigm, engineers for the software would go beyond the use of CNNs, RNNs, and gesture datasets to architect precise movement and articulation and would ensure that human capacity is not degraded from constant use of the tool. This could be done through an occasional quiz that helps to correct human gestures or translation such that the application does not cause human dependencies. The developers would explore ways to measure AI subordination such that human users have input into the evolution of the tool and its intentions. In a refined paradigm that elevates human empowerment, the cadre of deployers for any AI application would be most concerned with informing humans, not only of tool functionality, but of all that is necessary to gauge human agency and autonomy while using the tool. We can transform the research, design, development, and deployment of AI by guarding against over-reliance on such tools and by protecting and prioritizing the choices and access rights of the humans for whom such tools are developed.

Background

In recent years, there have been extraordinary contributions to the study of and practical outputs from AI that emphasize its impact on humans. To name a few, there are the Asilomar AI Principles (Asilomar AI Principles, 2017), a set of AI governance rules developed by a collection of industry, academic, and policy professionals; Human-Centered AI, an entire field of study with a focus on human well being, responsible design, and independent oversight; and the Blueprint for the AI Bill of Rights (OSTP, 2022), an extensive body of human-focused work from the White House Office of Science and Technology Policy. Even with these meaningful outputs, the AI research community's collective approach to appraising AI holds systemic misconceptions that impede our ability to research, design, develop, and deploy AI systems that empower humans. There is an underlying fallacy to which even the most "human-benefitting" methods of exploration fall prey. The fallacy is that we can somehow solve for human empowerment through the governance of a machine or algorithm and the measurement of its adherence to ethics, transparency, or safety rules and its alignment with human intentions. At best, testing for the presence of these characteristics can validate a system's human compatibility or alignment with human values. For this reason, the status quo human-centric AI practice as it currently exists is insufficient on its own.

This is not to say that the use of safety cases (Buhl et. al., 2024) as meaningful decision points in the development and deployment of large language models is not helpful or that there are not immense benefits in the conduct of model evaluations (Shevlane et. al., 2023). Approaches that include standards, concepts, and principles such as system fairness, accountability, beneficence, justice, explicability (Bocklisch and Huchler, 2023), interpretability (Veale et. al. 2018), and consentability (Kim, 2019) all offer meaningful information about an AI system's alignment with human values. However, although such constructs are worthy foci of model assessment, their collective framing is limited and, as noted in (Bowman and Dahl, 2021), it is not enough to simply create benchmarks that "disincentivize the use of systems with potentially harmful biases" (p. 2) since many benchmarks have "risks inherent in their framing" (Raji et. al., 2021, p.2). The aforementioned best-practices still center the machine, model, or algorithm and its various metrics in an effort to make safety determinations for humans. A more human-centered problem formation would involve a systemic adjustment of our tests and assessments of models such that the focus is on an outcome of human empowerment or the best possible future for humans in addition to focusing on model, machine, and algorithm acceptance or the least harmful future for humans. Our preference should extend to full subordination of AI under humans as a normative criteria (Wilkens et. al., 2021) toward human empowerment. Once established, such a standard could lead to the organization and regulation of human empowering AI.

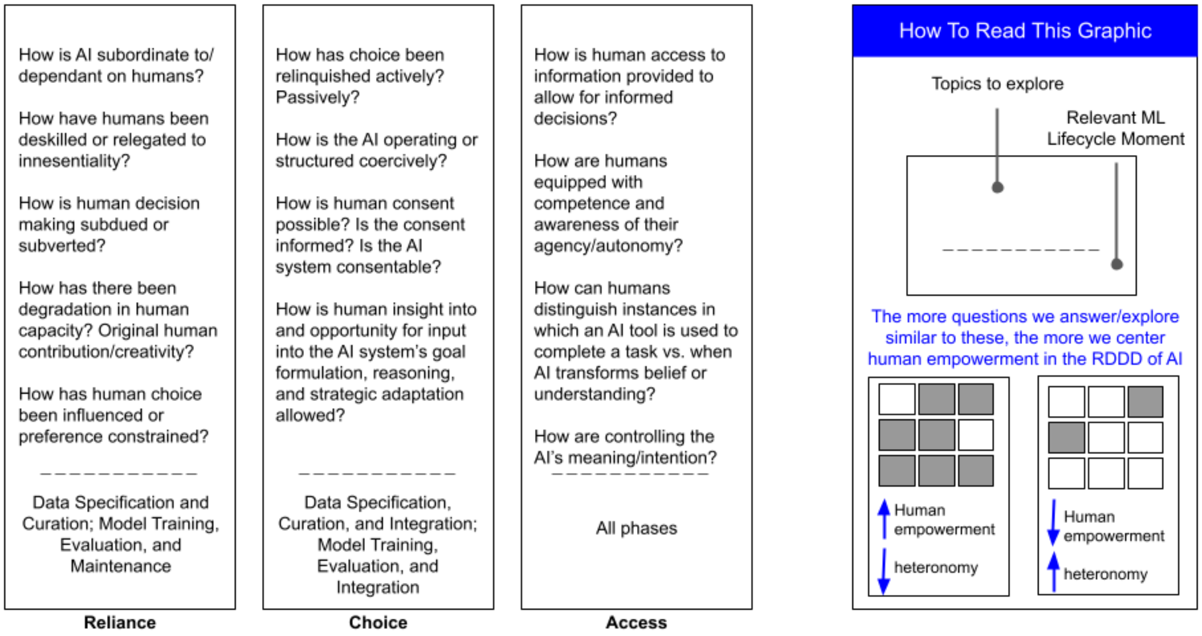
Diagnostic framework for human empowerment

This paper proposes topics and questions to explore in order for us to adjust our aim as we center humans in the RDDD of AI. Although the proposal does not include all possible topics and questions and although the vectors in use (e.g. reliance, access, and choice) are not the only anchors from which to explore human empowerment, they are meant to shift our thinking and focus as we aim for more than just a consideration of human values and compatibility. This AI Governance Workshop paper offers a starting point for human empowerment construct research. To generate a meaningful human empowerment paradigm, examples, case studies, and metrics are needed in order to demonstrate its operationalization. An entirely new methodology is needed.

Figure 1 is a pictorial representation of a human empowerment paradigm with questions and topics that are human empowerment-centered as opposed to machine, model, or algorithm focused. This paper posits that an exploration of human empowerment topics relevant to human reliance on AI, human choice, and human access are meaningful additions to the current approaches that include standards, concepts, and principles related to fairness, accountability, and safety. There is a difference between benign reliance and a disempowering over-reliance on AI. For this reason, we are less concerned with how often a human uses AI or for which tasks. In terms of reliance, the proposed framework endeavors to protect against the extensive deskilling or inessentiality, the constrained decision-making, and the degradation in capacity of humans. Taken together, the human empowerment paradigm and human-values-aligned parameters can support the investigation of measures that give insight into how aligned a system is with human intent and can explore human empowerment for an altogether more human-centered approach to AI model assessment. The idea is that human empowerment can be investigated and, with the development of future tools, quantified by the pursuit of answers to the questions related to reliance, choice, and access. If we pay attention to a majority of these categories, the technologies we build lean in the direction of human empowerment and we diminish heteronomy while an omission of such considerations will further diminish human empowerment and strengthen heteronomy.

Theoretical grounding

This paper builds on human-centered AI (HCAI), a set of standards, concepts, and principles such as fairness, accountability, beneficence, justice, and explicability (Bocklisch and Huchler, 2023). These and many other metrics including interpretability (Veale et. al. 2018), consentability



How is AI subordinate to/ dependant on humans?

How have humans been deskilled or relegated to innesentiality?

How is human decision making subdued or subverted?

How has there been degradation in human capacity? Original human contribution/creativity?

How has human choice been influenced or preference constrained?

Data Specification and Curation; Model Training, Evaluation, and Maintenance

How has choice been relinquished actively? Passively?

How is the AI operating or structured coercively?

How is human consent possible? Is the consent informed? Is the AI system consentable?

How is human insight into and opportunity for input into the AI system's goal formulation, reasoning, and strategic adaptation allowed?

Data Specification, Curation, and Integration; Model Training, Evaluation, and Integration

How is human access to information provided to allow for informed decisions?

How are humans equipped with competence and awareness of their agency/autonomy?

How can humans distinguish instances in which an AI tool is used to complete a task vs. when AI transforms belief or understanding?

How are controlling the AI's meaning/intention?

All phases

The more questions we answer/explore similar to these, the more we center human empowerment in the RDDD of AI

Figure 1: Diagnostic Framework: Empowering humans and reversing heteronomy - proposed questions and relevant topics for RDDD of AI throughout the ML lifecycle

(Kim, 2019), and generalizability (PAI, AboutML), to name a few, represent an approach to categorizing the extent to which a system is poised to consider humans. Taken together, these can be coalesced into a collection of measures that give insight into how aligned a system is with human intent.

However, how we test, evaluate, or appraise these notions still centers the machine, algorithm, or system and its various specifications making the technology the subject and its impact on humans as objects instead of centering human empowerment explicitly. A review of such concepts across many spheres of study can help us to use a human-centricity theory to bridge the gap between technology and societal deployment or application. This paper endeavors to express the shift in framing as a radical reimagining in which the measurement of machine or algorithm adherence to scales of ethical compliance are an incomplete means to centering humans in the RDDD of AI systems. Kluge et. al. (2024) note that "current research streams are dominated by techno-centric and engineering perspectives" (p. 1) further solidifying the need for this work to consider structural risks and how AI interacts with social, political, economic forces in society and other facets beyond AI (Zwetsloot and Dafoe, 2019).

Concept Discussion

This section offers insights for how we currently understand challenges AI poses to human autonomy agency.

Autonomy According to Prunkl (2024, p. 25) "autonomy broadly refers to a person's effective capacity to act on the basis of beliefs, values, motivations, and reasons that are in some important sense her own". This is a concept that is difficult to study and design algorithmically since it is so context dependent and multifaceted and since it is not clear whether the beliefs in question are authentic, distorted, or manipulated. Who is to say an authentic sense of autonomy is not externally impacted? Autonomy is challenging to measure since we are all products of our environment which include externalities and some of the research exploring human autonomy as it relates to AI systems is set in the context of online manipulation (e.g. recommender systems or other technological mediations) which is built on the influence of a vast external digital environment. Another reason autonomy is tough to investigate is due to the architecture of AI systems that can influence choice (present tense) and also impact or constrain preference (future tense) (Adomavicius et. al., 2019).

For this paper, human autonomy in relation to AI is focused on *avoiding over-reliance* which encompasses deskilling, or the idea that the work left for humans after AI system activation will require a lower level of skill (Crowston et. al., 2024); technology dominance such that technology subverts and subdues human decision-making (Sutton et. al., 2023); and degraded human capacity and original authorship as factors. Creating a paradigm for human empowerment that helps to avoid over-reliance on AI is useful in managing the influence on and protecting the preferences of humans. Such a paradigm should be diligently constructed to include a localized adaptation of human empowerment that considers different social groups and societies since the "conditions that shape these concepts are fundamentally different in different regions" (Ewuoso, 2023 p. 3) of the world.

Agency Agency is presented by Prunkl (2024) as the external dimension of autonomy or a human's capacity to enact decisions, make choices, and take charge. Mackenzie (2014) likens agency to self determination and one of the Asilomar Principles (2017) states that "[h]umans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives". DiBlasi et. al. (2020, p.1) declare that "agency is how autonomy is exercised, articulated and maintained, via capacities such as adaptability, viability and sentience".

The intermingling of human agency and AI agency presents a slew of concerns including what Yeung (2017) considers coercive architectures which can limit consumer options (Prunkl, 2024) and constrain preferences (Chang and Cikara, 2018). In such cases, the agency of the AI system can degrade the freedom and opportunity conditions (Mackenzie 2014) of the agency of the human. Action-forcing or the coercive design of system architecture automates the decision process and is different from decision-guidance architecture where the human is making the decision based on optimality and suggestions (selection optimization) (Yeung, 2017). Covert and manipulative uses of big data can influence behavior further exploiting cognitive irrationalities. Manipulated results benefit the gatekeeper and distort the capacity of individuals to make informed, meaningful choices (undermining individual agency).

For the purposes of this paper, human agency will be the focus and will center around *human choice* which can be relinquished both actively and passively and is often obfuscated via coercive architectures. Creating a paradigm for human empowerment that helps to preserve human choice is useful in protecting and promoting human consent. Such a paradigm could transform the participation of humans such that goal formulation, logical reasoning, and strategic adaptation of the AI system involve the user by default. The paradigm should be thoughtfully architected such that a "structural decolonisation … that seeks to undo colonial mechanisms of power, economics, language, culture and thinking that shapes contemporary life" can be conducted and more than just the "dominant forms of knowledge, values, norms and assumptions" (Mohamed, et. al., 2020) become embedded in the evaluation of an AI system.

Disembodiment Embodied AI can refer to the idea that an AI algorithm and agents no longer solely learn from curated images, videos, or text but through interactions with environments from an ecocentric perception similar to humans (Duan et. al, 2022). It is possible that the achievement of a fully embodied AI system could simultaneously influence the disembodiment of humans. The meaning-making and computational work of an AI system can impact our beliefs as well as our knowledge. This exalts an AI system from a tool we use to complete a task to a technology that wholly changes what we believe about our world. What might this mean when considering an AI system's impact on our knowledge and/or cognition? Dennett (1993) notes that our self understanding presupposes cognitive notions such as believing, desiring, and knowing but does not explain them. Can an AI system impact our self understanding and further explain or make meaning for us (whether that meaning is true or not)? This paper posits that humans can become disembodied from our own understanding of self and embodied in that of a meaning-making AI system leading to a state of human disempowerment.

Humans can be disempowered and AI has power over us if a system's programming derives from its own learning and operates at speed and scale beyond the effective control of those who design and deploy it. It can be said that the control is embodied in the AI system. Even small individual effects can translate into significant aggregates intervening on human interests and shaping our options and beliefs and desires or even ourselves. This could represent AI technology's commandeering of human agency and autonomy. How do we maintain our sense of self, our ability to think and choose for ourselves, for example when we are constantly entangled with AI systems that are designed to learn from us and influence us, often without our full understanding or control? Can we remain embodied in our own agency through the often dispossessing effects of AI system use?

For the purposes of this paper, disembodiment will be showcased as a measure of both *access and control*. The concept of access is a broad one that includes the ability to obtain information necessary to make informed decisions (e.g. insight into weights for open models, access to training data, knowledge of AI system mediation in transactions, self understanding, etc.). Large language models are said to be transforming our economic and political lives (Lazar, 2024) and we are increasingly subject to power/control exercised by means of automated systems which underpin vital government services and determine how we find out information. Creating a paradigm for human empowerment that helps to prioritize human access and control can be helpful as humans make informed decisions about the use of AI. Such a paradigm could encourage a competent human awareness of agency and autonomy such that tool use is not inadvertently morphed into a transformation of belief and understanding for humans.

Heteronomy This paper is primarily concerned with the way humans can be empowered through the responsible RDDD of AI and how the concept of heteronomy can be diminished if not avoided through this responsible, human empowerment-centered focus.

According to Sending (2016, p. 10) "Heteronomy is the opposite of autonomy: it is not self-rule, but a condition of rule of other, or others, which is indirect, and often not recognised as such." In relation to how humans interact with AI, heteronomy is an important concept to investigate as we endeavor to avoid the unconscious and conscious relinquishing of decisions, influence, and actions while we evolve our use of AI. Handing over such control requires us to commit to the idea that "AI knows best" and resigns us to a system that chooses "roles and obligations" (Onuf, 2012) for us as human actors. Heteronomy, used here as a state in which AI systems control human autonomy and agency, is often disguised as instruction or direction or even evidence-based information to be used as input to human decision-making. It has the illusion of independence (Onuf and Klink, 1989), yet one of the distinguishing features of heteronomy is that it is almost impossible to identify the authors or originators of the rules that we as humans are following. Is the developer of the system influencing the direction we take? Is the system's programming, derived from its own learning, mediating human action? Formosa (2021, p. 602) declares that "[e]ven if humans remain formally part of the decision loop, they may be biased towards always uncritically following the machine's advice, which practically means that they are allowing the machine to act with little or no human oversight (as in a self-imposed "no way out" design)." In this sense, heteronomy marks a dangerous shift in our relating with AI systems such that human-centeredness is not the primary outcome of the endeavor.

For this paper, human autonomy is viewed as avoiding over-reliance on AI, agency as human choice, and disembodiment as a measure of human access and control. Together these components create the foundation of an ecocentric lens into heteronomy; the rise of which can impede human empowerment. In our efforts to create conditions of possibility for humans to survive technology, a thoughtful restraint allowing humans to keep pace with what we have created and to bring AI into full submission under human empowerment could more than benefit society. It could save humanity.

Future work and recommendations

This AI Governance Workshop paper attempted to provide visionary perspectives to guide future research and, although it explored a meaning for the concept of human empowerment, the capacity to measure and govern this idea has not been developed. Out of scope for this work is the concept of human empowerment enforcement, the complexities of our current geopolitical landscape of power and governance, and our historical allegiance to capitalism over humanity that is showcased by algorithms that prioritize profits over truth (Noble, 2018).

Future explorations into the empirical analysis for AI assessment as well as outcome measures and case illustrations to further test the ideas of a human empowerment paradigm could be meaningful. To generate a human empowerment paradigm, examples, case studies, and metrics are needed in order to demonstrate its operationalization. An entirely new methodology is needed. Ways to quantify the protection of human influence, to estimate the degree of disempowerment, and to better characterize civilization-scale multi-agent dynamics (Kulveit et. al., 2025) could be meaningful extensions of this work. Furthering the idea from Prunkl (2024, p. 6) that explains how "being explicit about the requirements for autonomy can help us to distill conditions that need to be fulfilled for a system to count as autonomy-promoting or autonomy-undermining" seems like a relevant exploration. Finding ways to operationalize and test a human empowerment paradigm is one way to resist what Carrigan et al. 2021 call "techniques of invisibility," the means by which computing corporations "clos[e] down the possibility for transparency in the production and applications of machine learning algorithms. . . and help exempt Big Tech from ethical standards enforced in society, for example, protections for privacy and bodily autonomy" (p. 2). It would be a meaningful progression in the ethical AI field if we could create viable learning problems or benchmarks that were developed as abstractions of the human empowerment paradigm and systematically curated in the realm of that construct (Raji et. al., 2021). Additionally, future research should give thought to the dynamic nature of how humans are shaped by AI which would add complexity to the human empowerment construct. There must be a trade-off discussion that considers human empowerment against traditional ideas of AI scalability, efficiency, and commercial viability in the development of this new paradigm.

Conclusions

If, in the near future, artificial agents are to serve human needs (Chakraborty and Bhuyan, 2024), shouldn't we be highly in tune with what those needs are? This work is an imperfect mirror to Johan Galtung's theoretical concept of "positive peace, which goes beyond the mere absence of war (negative peace) to include the presence of social justice and the integration of human society" (VOH, 2024). The argument is that, in order to properly center humans in the RDDD of AI, we require a focus on an outcome of human empowerment (or the best possible future for humans) in addition to our current focus on model, machine, and algorithm safety acceptance (or the least harmful future for humans). The paper proposes a structure and diagnostic framework for defining human empowerment that is a precursor to an objective and universalizable methodology.

Acknowledgments

With appreciation to Coleen Carrigan, Copenhaver Associate Professor at UVA, for her in-depth review and feedback. Dr. Carrigan introduced a possibility for future work based on the explicit exploration of research questions such as "Who is disempowering humans and centering machines in the production and distribution of AI? It is not the machines themselves so what humans need to be challenged in the current power relations under which AI is produced to reform these machines to better serve humanity?" StoryTherapy LLC. provided a meaningful review for clarity to a general audience. The affiliates of the Science, Technology, and Social Values (STSV) Lab attended a focus group discussion to provide meaningful and directional feedback for the work. Affiliates include Ulla Berg, Tatiana Carayannis, Hannah Bloch-Wehba, Ellie M. Hisama, Brian Jordan Jefferson, Nahoko Kameo, Nicholas Collins, and Marie-Therese Png.

The author is affiliated with the Institute for Advanced Study and Dr. Alondra Nelson's STSV Lab as well as participation in the AI Policy and Governance Working Group. The author's primary funding for this work came from the Kavli Foundation and the Rita Allen Foundation.

References

Bocklisch, F., and Huchler, N. (2023). Humans and cyber-physical systems as teammates? Characteristics and applicability of the human-machine-teaming concept in intelligent manufacturing. Frontiers in artificial intelligence, 6, 1247755.

Buhl, M. D., Sett, G., Koessler, L., Schuett, J., and Anderljung, M. (2024). Safety cases for frontier AI. arXiv preprint arXiv:2410.21572.

Carrigan, C., Green, M. W., and Rahman-Davies, A. (2021). "The revolution will not be supervised": Consent and open secrets in data science. Big data and society, 8(2), 20539517211035673.

Chakraborty, A., and Bhuyan, N. (2024). Can artificial intelligence be a Kantian moral agent? On moral autonomy of AI system. AI and Ethics, 4(2), 325-331.

Chang, L. W., and Cikara, M. (2018). Social decoys: Leveraging choice architecture to alter social preferences. Journal of personality and social psychology, 115(2), 206.

Dennett, D. C. (1993). The embodied mind: Cognitive science and human experience.

DiBlasi, J., Castellanos, C., Kang, E., Poltronieri, F., and Smith, L. (2020, November). Agency and autonomy: Intersections of artificial intelligence and creative practice. In International Symposium on Electronic Art (Vol. 7).

Duan, J., Yu, S., Tan, H. L., Zhu, H., and Tan, C. (2022). A survey of embodied ai: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(2), 230-244.

Ewuoso, C. (2023). Black box problem and African views of trust. Humanities and Social Sciences Communications, 10(1), 1–11.

Formosa, P. (2021). Robot autonomy vs. human autonomy: social robots, artificial intelligence (AI), and the nature of autonomy. Minds and Machines, 31(4), 595-616.

Kim, N. S. (2019). Consentability: Consent and its limits. Cambridge University Press.

Kluge, A., Wilkens, U., Nitsch, V., and Peifer, C. (2024). Human-centered AI at work: common ground in theories and methods. Frontiers in Artificial Intelligence, 7, 1411795.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. (2025). Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. arXiv preprint arXiv:2501.16946.

Lazar, S. (2024). Automatic Authorities: Power and AI. arXiv preprint arXiv:2404.05990.

Mackenzie, C. (2014). Three dimensions of autonomy: A relational analysis. Oxford University Press.

Mohamed, S., Png, MT. and Isaac, W. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philos. Technol. 33, 659–684 (2020).

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In Algorithms of oppression. New York university press.

Onuf, N., and Klink, F. F. (1989). Anarchy, authority, rule. International Studies Quarterly, 33(2), 149-173.

Partnership on AI, Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (AboutML), https://partnershiponai.org/workstream/about-ml/

Prunkl, C. (2024). Human autonomy at risk? An analysis of the challenges from AI. Minds and Machines, 34(3), 26.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). AI and the everything in the whole wide world benchmark.

Sending, O. J. (2016). Agency, order, and heteronomy. European Review of International Studies, 3(3), 63-75.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... and Dafoe, A. (2023). Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

Veale, M., Van Kleek, M., and Binns, R. (2018, April). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14).

Vision of Humanity, (February 19, 2024), Johan Galtung, 1930 – 2024: A Life Dedicated to Peace

Wilkens, U., Reyes, C. C., Treude, T., and Kluge, A. (2021). Understandings and perspectives of human-centered AI–A transdisciplinary literature review. GfA Frühjahrskongress, B, 10.

Wired. (2025, May 8), Hill, Simon. This startup has created AI-powered signing avatars for the deaf. https://www.wired.com/story/silence-speaks-deaf-ai-signing

The White House, Office of Science and Technology Policy. (2022). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

Yeung, K., (2017) Hypernudge': big Data as a mode of regulation by design Inf. Commun. Soc., 20 (1), pp. 118-136

Zwetsloot, R., and Dafoe, A. (2019). Thinking about risks from AI: accidents, misuse and structure. Lawfare.