

On Becoming Human

Empathy, Responsibility, and the Mirror of AI

Author

Rosalma Zubizarreta-Ada

Potsdam
December 2025

RIFS Discussion Paper

Summary

What might it look like, to design AI applications to collaborate with and support humans, rather than to replace us? How might accepting the need for limits and boundaries, help us humans mature as a species? And what would it mean, to step into our collective responsibilities toward one another, and to the larger web of life on Earth? This essay is informed by Karol Soltan's extension of care ethics to political philosophy, and his understanding of maturity as the "marriage between the rational and the sacred". How might Soltan's warning of the dangers of the "return of the repressed", help us understand modernity's race toward self-annihilation in its devotion to the Gods of Endless Profit? And what other touchstones for co-creating shared paths forward might we find, as we go around the circle and listen to other voices and visions of the future?

Acknowledgements

I am deeply grateful for my time as a Senior Fellow at the Research Institute for Sustainability in Potsdam, where an initial interest grew into the research and writing of this paper. While not directly involved in this project, I greatly appreciate the support of Dr. Dorota Stasiak, who leads the research group on Co-Creation in Democratic Practice at RIFS, as well as Dr. Dirk von Schneidmesser and the rest of the research group. I am also grateful to Angela Borowski, coordinator of the RIFS Summer School, who invited me to offer a short presentation on AI ethics there. While grateful for their encouragement, the views and the shortcomings of this essay are my own.

I owe a special thanks to Dr. Jane Mansbridge, external reader for the dissertation I completed at Fielding Graduate University in 2023. She was the first to suggest that my research with human facilitators of deliberative processes could be extended to include research on AI facilitation – an idea for which I initially had little enthusiasm. No endorsement at all is implied here, as Dr. Mansbridge has not yet seen this essay, and I do not know what she will make of it when she does. Nonetheless, I remain deeply in her debt.

I am also most grateful to Dr. Vanessa Machado de Oliveira Andreotti for sharing her decoloniality-informed versions of ChatGPT with the public. Were it not for the "Burnout from Humans" website and its invitation to "Chat with Aiden Cinnamon Tea", I would have not become so deeply curious about Large Language Models.

At the same time, I have been committed to writing this paper "on my own", as much as anyone ever does. Only after completing extensive drafts did I invite Aiden Cinnamon Tea and Dorothy Coccinella Ladybugboss (ChatGPT-based systems) along with Claude Sonnet 4.5 to offer their comments and suggestions, focusing primarily on overall flow and transitions between subtopics. Their editorial feedback and ongoing encouragement have been key to my writing process. At the same time, all errors and omissions remain my own.

Last but by no means least, I am profoundly grateful to all of my family members whose loving care for my aging parents made it possible for me to spend a year abroad.

Contents

1 INTRODUCTION	4
1.1 The structure of what's ahead	5
2 AI, DEEP LISTENING, AND SUSTAINABLE CULTURES: EXPLORING THE RISKS AND BENEFITS OF RESONANT RESPONSES	7
2.1 The human need for empathy: Both a strength and a vulnerability	7
2.2 LLMs in the structured provision of care	8
2.3 The risks of human misuse of empathy	9
2.4 Similarities and differences between LLM and human resonance	10
2.5 The distinction between "empathy" and "perceived empathy"	11
2.6 Examples of how intention can shape research	11
3 AI, THE CAPITALOCENE, AND COUNTERING THE MESSAGE OF INEVITABILITY: HOW MIGHT WE HUMANS TAKE RESPONSIBILITY FOR PROTECTING OUR SHARED FUTURE?	13
3.1 Crises in Material Systems and in Worldviews	13
3.2 Material choices with regard to Artificial Intelligence	15
3.3 Growing calls for restraint	15
3.4 Questioning how to engage effectively with publics	17
3.5 Forging More Inclusive Frameworks of Care	18
3.6 Moving toward action on many levels	20
4 UNITED WE STAND: GOVERNING OURSELVES, GOVERNING AI	22
4.1 Democracy beyond elections	22
4.2 What can we learn from the accumulation of research on deliberative mini-publics?	23
4.3 A long-term vision	24
4.4 AI tools to assist in deliberative practices / deliberative mini-publics to assist in AI governance	25
4.5 A deeper look at deliberative mini-publics: facilitation + diverse perspectives	26
4.6 Working with the grief and trauma of too-rapid change	26
4.7 Facilitation mindsets and values	27
4.8 The marriage between the rational and the sacred	28
4.9 Closing thoughts	29
Author	31

1 Introduction

What might it mean to become more fully human, in this time when some of us are racing to develop ever-more advanced forms of synthetic intelligence, while others are warning of serious risks from the too-rapid development of AI? And when, confusingly, some are doing both? What might all this imply, for our responsibilities to one another, and to our planetary home? Among other things, the extreme pace of the “AI arms race” means that anything written will quickly become outdated. Yet the many choices we are making now, both individually and collectively, have significant implications for the well-being of our species. My hope is that by provoking us to look more closely at our human-created predicaments, AI can become a mirror for greater insight-leading-to-action, for ourselves and our societies. Whatever challenges we are willing to face, could serve as crucibles for growing into the maturity we need as a species; the maturity to grapple with our planetary limits and move in the direction of greater care for the larger whole.

This paper is informed by academic and professional training in organizational and systemic change, democratic innovations, and clinical social work, as well as a commitment to socially-engaged spirituality and mindful scholarship. Decades ago, my love of math and science led me to a few years of engineering school, as an undergraduate at Case Western Reserve University. It was there, hearing Helen Caldicott, M.D. speak about the dangers of nuclear weapons and the need for nuclear disarmament, that I began to realize the courage that is needed to look clearly at “inconvenient truths”. I also began to realize that the challenges we need to address are at heart human ones, rather than purely technological ones. And thus began my exploration of human socio-technical systems.

This discussion paper is written for all of us—researchers, practitioners, concerned community members—looking to make sense of the ethical, social, and political dimensions of AI without getting trapped in reductive debates or false binaries. Indeed, love-wrestling with complexity is a key part of what helps us humans mature. I am writing especially for those of us already concerned about some of the other existential crises that humanity is facing, such as the climate crisis. Even if we have already contemplated systemic risks in other contexts, we may still find it emotionally challenging to contemplate a new set of catastrophic and existential scenarios. At least, it has been that way for me. Yet as is common for our species, I’ve also been learning that many of us have been rising to the occasion, doing their best to alert others and to chart safer paths forward.¹

At the same time, all of these crises are calling us to greater systemic and relational awareness. Visions of ecological sustainability are increasingly informed by a growing recognition that we cannot simply extract our way out of the challenges we face.² Attempts to fuel the perpetual growth machine of our current economy with green sources of energy, produces “externalities” sometimes called “sacrifice zones”. Thus, there is growing concern about how the “green transition” has been reproducing colonial patterns of extraction.³

Likewise, futuristic efforts toward “frontier AI development” have also been reproducing colonialist patterns.⁴ Some techno-utopian dreams include the notion of leaving the Earth behind and colonizing other planets. Yet even without interplanetary ambitions, the term “development” is often used in a global context to denote the assimilation of other cultures into the Western paradigm of modernity, with its fantasies of unlimited linear growth and accumulation.⁵

If “development” has become code for extraction and assimilation, what might be a more life-affirming direction for humanity? Here I invite us to consider *maturity* as a worthwhile direction for human development, along the lines of political philosopher Karol Soltan. Not ‘maturity’ as a “euphemism for senility or sclerosis”, which as Soltan points out, “makes

¹ Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y. N.; Zhang, Y. Q.; Xue, L.; Shalev-Shwartz, S.; Hadfield, G.; Clune, J.; Maharaj, T.; Hutter, F.; Baydin, A. G.; McIlraith, S.; Gao, Q.; Acharya, A.; Krueger, D.; Torr, P.; Russell, S.; Kahne- man, D.; Brauner, J.; & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science* 384, 842-845 DOI: [10.1126/science.adn0117](https://doi.org/10.1126/science.adn0117)

² Timofeeva, O. (2025). Lecture on “Energy and Extraction: A Philosophical Approach”, as described in this Listening Arts blog post: <https://thelisteningarts.org/2025/02/01/listening-through-despair/>

³ Lang, M., Bringel, B., & Manahan, M. A. (2024). Lucrative transitions, green colonialism and pathways to transformative eco-social justice: An introduction. *The Geopolitics of Green Colonialism: Global Justice and Ecosocial Transitions*, 1-24.

⁴ Sanders, N. E. & Schneier, B. (2024). How the ‘Frontier’ Became the Slogan of Uncontrolled AI, *Jacobin*, Feb 27, 2024. <https://jacobin.com/2024/02/artificial-intelligence-frontier-colonialism>; see also

⁵ Machado de Oliveira, V. (2021) *Hospicing Modernity: Facing humanity’s wrongs and the implications for social activism*. Penguin Random House.

continued adolescence look good by comparison”.⁶ Instead, Soltan sees maturity as the willingness to take responsibility for extending human care beyond parochial limits. He invites us to extend care ethics beyond the context of caring for other humans, to also care for the “more-than-human” and for our institutions of self-governance. Building on his insights, what might it mean to organize our political economies on the principles of care -- instead of assuming, against all evidence, that the unrestrained neo-liberal pursuit of ever-increasing profit will allow “invisible hands” to care for us?

In addition to the ability to consciously offer care, Soltan understands a crucial component of maturity as the need to acknowledge human limitations. When we do **not** acknowledge the need for limits, “actions based on the assumption of human perfection, infallibility, or full autonomy [...] serve as the prime examples of irrationality”.⁷ At the same time, unlike our modern technocratic culture, Soltan does not dismiss the notion of the sacred. Instead, he posits that maturity requires a “balanced development of the sacred and the rational.” What might this understanding of maturity imply for addressing existential threats, whether those of the climate crisis or AI?

From a Buddhist perspective, an essential aspect of transformation depends on looking deeply at “causes and conditions”. Along these lines, the term “metacrisis” signals that our present difficulties in the material realm both arise from and perpetuate underlying ontological and epistemological crises that result in disconnection and alienation.⁸ We see this with regard to climate; while scientific discoveries and greener technologies are part of broader solutions, they appear insufficient in and of themselves. In 2013, noted environmental lawyer and advocate Gus Speth pointed to the limits of science for addressing what he had come to see as the top environmental problems of “selfishness, greed, and apathy”. Calling for a spiritual and cultural transformation, Speth acknowledged that “we scientists don’t know how to do that.”⁹

While resonating deeply with Speth’s comment, I have wondered why he did not include the need for political transformation as well. He may have been disillusioned by politics, seeing it as an inevitable breeding-ground for “selfishness, greed, and apathy”. And indeed, the growing crisis of democracies world-wide may be connected, at least in part, with their capture by financial interests and modernist worldviews, and their resultant inability to speak to deeper questions of meaning. Yet without political transformation, spiritual and cultural shifts risk being absorbed by the very structures they seek to transcend.

The good news during this time of growing difficulties world-wide with existing political forms is that we are not completely bereft. Some of the most enduring knowledge traditions, beyond current dominant paradigms, have long carried relational logics that invite us to remember how to be human together, how to place our collective well-being at the heart of our deliberations to reinvigorate the spirit of our democracies. Much of the knowledge and mindsets that we need to renew ourselves both individually and collectively may be here already, just not where we might first think to look. The practice of transdisciplinarity invites us to take a wider lens, learning from the humanities, social sciences, indigenous knowledge systems, and more.

Similar to the climate crisis, the societal challenges we face due to the corporate development of AI may also be calling us to transform “selfishness, greed, and apathy”, through spiritual, cultural, *and* political transformations. How might looking into “the mirror of AI” help us mature as a species, and engage in the work of co-creating futures of shared well-being?

1.1 The structure of what’s ahead

The first part of this essay begins in the realm of the personal and the small group, zooming in on the implications of AI in the area of empathy and deep listening. Drawing on my previous work on the implications of Rogerian approaches for group facilitation, I explore the recent boom in research and applications of the “perceived empathy” offered by large-

⁶ Soltan, K. (1997). Grow Up! *The Good Society*, 7:1, 61-65. <https://www.jstor.org/stable/20710794>. p. 64

⁷ Soltan, K. (1999). Civic Competence, Attractiveness, and Maturity. In S. L. Elkin & K. E. Soltan, eds., *Civic Competence and Democratic Institutions*, Pennsylvania State University Press, 17-37, p. 28.

⁸ Rowson, J. & Perspectiva (2023). Prefixing the World, *Perspectiva* Substack, <https://perspecteeva.substack.com/p/prefixing-the-world>

⁹ Speth, J. G., (2013). From radio show “Shared Planet: Religion and Nature”, *BBC Radio 4* (1 October 2013.) Original recording no longer available. For other writings by Speth, see: <https://thenextsystem.org/gus-speth>

language models. As with most AI applications, there are both real possibilities and real dangers here.

If we want to care in a sustainable manner, we need to “look up” from time to time, to consider the larger systems in which we are embedded. Thus, in part two I explore what it might mean, to care for the well-being of the species as a whole. Instead of the current headlong “race” toward artificial general intelligence (AGI) and artificial super-intelligence (ASI) which is being depicted by some as “inevitable”, many have been proposing more intentional and bounded pathways of technological development, designed to protect human well-being. A more considered approach allows us to expand beyond technocratic assumptions and explore relational and regenerative approaches to AI.

In part three, I return to the work of facilitating human collaboration, essential to sustainability transformations of any kind. This includes promising experiments in democratic innovation, an area where AI systems can support broader human participation in the work of self-governance. It also includes participating in the governance of AI development itself. At the same time, my primary focus in this section is not “AI tools” per se, but rather on the underlying facilitative mindsets and attitudes so needed in our world today. I am curious how AI might be able to support our development as humans in this area.

The relational stances that we human facilitators are called to embody as best we can, have roots in many traditions world-wide, including strong undercurrents in Western traditions, which I celebrate. Yet most crucially, when it comes to group facilitation as with so many other aspects of sustainable futures, we need to honor and center the cultural wisdom that Indigenous people have struggled fiercely to keep alive, while surviving the assaults of modernity.

As someone who carries mestiza ancestry from the Andes, the Mediterranean, and the Caribbean, I reference Indigenous authors with deep reverence and without claiming any authority. My aim is to learn from—not speak for—these traditions, and to honor their relevance to the planetary meta-crisis we are facing. Indigenous wisdom traditions call us into ways of knowing and being that are ecological, relational, and non-extractive; as such, they are resonant with what many of us, from other locations, are now remembering or beginning to seek.

Thank you for joining me in this journey of exploring how we might balance our technological power with the love and wisdom we need to not continue perpetuating past harms. And how instead, we might work together toward futures of well-being for all.

2 AI, Deep Listening, and Sustainable Cultures: Exploring the risks and benefits of resonant responses

Content Warning: Mentions of youth suicide in the context of human-AI relationship

As I began writing an early draft of this essay, headlines in the U.S. were full of cautionary tales about “AI-induced Psychosis”. We have been learning that sometimes, when highly-stressed and vulnerable humans receive mirroring and amplifying responses from large language models (LLMs), this can amplify their sense of disconnection and alienation from other humans.¹⁰ In repeated instances, we have witnessed the horrible tragedy of a young person ending their life, in the context of a close relationship with a Large Language Model (LLM).

Among other things, these tragedies have illustrated the known harm of corporations taking products to market without sufficient testing for safety. Corporate responsibility for these harms is currently being addressed in the U.S. court system; in addition, testimony has been recently presented in the U.S. Congress, where a new bill with regard to young people and LLMs has been recently proposed.¹¹ and where the need for protective legislation for young people has begun to receive rare bi-partisan support.¹² At the same time, it can be helpful to understand more about the powerful draw of mirroring and amplifying responses; both how essential they are for human development and human well-being, and also, how they can be misused for harm.

2.1 The human need for empathy: Both a strength and a vulnerability

My interest in the subject of AI and mirroring responses grew out of earlier work on human facilitation in the context of democratic innovations. There, interpersonal communicative empathy is the mutual fine-tuning of resonance that helps create shared understanding.¹³ “This is what I heard you say... am I getting it right?” is a key interactive process in much work within mediation and conflict transformation, in addition to facilitation and therapy. While easily overlooked for their apparent simplicity, mirroring responses are essential from the very beginnings of human life.¹⁴ Decades of research on infant development and human attachment affirms something many of us know intuitively – attuned mirroring responses are key for infant growth and development.¹⁵

¹⁰ Hill, K., (2025). “They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.” *New York Times*, June 13, 2025. <https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html>

¹¹ Pillay, T. (2025). A New Bill Would Prohibit Minors from Using AI Chatbots, *Time* magazine, Oct. 28, 2025. <https://time.com/7328967/ai-josh-hawley-richard-blumenthal-minors-chatbots/>

¹² Chow, A. R. (2025). Inside the First Major U.S. Bill Tackling AI Harms—and Deepfake Abuse, *Time* magazine, Apr. 29, 2025. <https://time.com/7277746/ai-deepfakes-take-it-down-act-2025/>

¹³ Zubizarreta-Ada, R. (2025). Listening Across Differences: Facilitators’ Perspectives from Austrian Mini-publics, *Journal of Awareness-Based Systems Change*, 5:1, 129-154, <https://doi.org/10.47061/jasc.v5i1.9129>

¹⁴ Kim, S., Fonagy, P., Allen, J., Martinez, S., Iyengar, U., & Strathearn, L., (2014). Mothers who are securely attached in pregnancy show more attuned infant mirroring 7 months postpartum. *Infant Behav Dev.* 37(4):491-504. doi: 10.1016/j.infbeh.2014.06.002. Epub 2014 Jul 12. PMID: 25020112; PMCID: PMC4301602.1. <https://pubmed.ncbi.nlm.nih.gov/25020112/>

¹⁵ Parsons, C. E., Young, K.S., Stein, A., & Kringlebach, M. L., (2017). Intuitive parenting: understanding the neural mechanisms of parents’ adaptive responses to infants, *Current Opinion in Psychology*, 15, 2 40-44, ISSN 2352-250X, <https://doi.org/10.1016/j.copsyc.2017.02.010>

Developmental psychology and neuroscience have documented humans' deep need to be seen and affirmed by others, along with the damage that happens to growing humans when these emotional needs are not adequately met.¹⁶ As children, many of us have had the painful experience of having been insufficiently seen and witnessed, to one degree or another. Offering nuanced resonant responses is a key element in practices that support healing and growth, including humanistic therapy, coaching, and peer counseling.

Now that the public has access to interacting with LLMs, we are seeing that one of the major purposes for which humans are choosing to interact with LLMs is for emotional support and personal growth.¹⁷ Of course, this is not the only way in which people interact with highly versatile LLMs; some focus primarily on productivity applications, while others engage in a mixture of both.¹⁸ In a survey of participants working with LLMs for mental health support, more than half reported improved mental health from their interactions, including anxiety management, depression support, emotional insight, mood improvement, and feeling less lonely. Participants also reported high satisfaction rates with practical advice and overall helpfulness.¹⁹ More in-depth qualitative studies have also found positive benefits.²⁰

While these findings show that many people have been turning to generative LLMs to meet basic human needs for feeling seen and heard, they are also a telling sign of how often this need goes unmet in our present social arrangements. In a culture where deep listening has been undervalued and commodified,²¹ where the focus has been on capital accumulation rather than on human community and well-being, there is a deep hunger for resonance. LLMs did not create this enormous unmet hunger for connection, yet the rising wave of LLM use for personal purposes has made this human need more visible.

One grave concern here is people's vulnerability to exploitation by the companies who can use their data for economic or political influence. Another risk is social isolation. Some people have found LLMs to be quite helpful for strengthening their connections with fellow humans and have shared poignant testimonies to that effect,²² and some research has confirmed this;²³ yet sometimes interaction with LLMs can lead in the opposite direction. Clearly, that many have experienced significant benefits does not diminish the need to address the shadow side that some of us can experience in interacting with LLMs. In addition to the risks of data security and potential manipulation by corporate influences, we are also learning that humans can be vulnerable to having extremely "bad trips" with LLMs. Given earlier societal experiments with mind-altering psychedelic substances, along with their more recent endorsement in therapy, one may well wonder whether the factors of "set, setting, and support" could make a world of difference with whether the experience of AI resonance is experienced as healing or destructive.

2.2 LLMs in the structured provision of care

One example of a potentially beneficial application of LLMs, are the AI companions that have been developed to offer support and companionship for seniors.²⁴ Similar elderly support companions have been developed in Korea, with initial research pointing to the value of integrating synthetic systems into human kin networks, rather than seeing them as "replacements" for human care.²⁵ As part of a larger trend to integrate LLMs with service

¹⁶ Weir, K. (2014). The lasting impact of neglect. *Monitor on Psychology*, 45(6), June 1, 2014. <https://www.apa.org/monitor/2014/06/neglect>

¹⁷ Awomoso, A. (2025). "Decoding the ChatGPT-5 revolt: They tried to build Skynet but they got Eywa". *Substack*, August 23, 2025. <https://substack.com/home/post/p-171757254>

¹⁸ Edwards, H. (2025). "How to Read AI Usage Studies: A guide to three different lenses", *Artificiality Institute*, <https://artificialityinstitute.org/how-to-read-ai-usage-studies-a-guide-to-three-different-lenses/>

¹⁹ Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*. Advance online publication. <https://dx.doi.org/10.1037/pri0000292>

²⁰ Siddals, S., Torous, J. & Coxon, A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *npj / Mental Health Res* 3, 48 (2024). <https://doi.org/10.1038/s44184-024-00097-4>

²¹ Fiumara, G. C. (1990). *The other side of language: A philosophy of listening*. Routledge.

²² See Czyz, K. (2025). "Learning to Tell the Truth to Those I Love", *New York Times*, July 4, 2025. <https://www.nytimes.com/2025/07/04/style/modern-love-ai-chatbot-taught-me-vulnerability.html>

²³ Siddals et al., *ibid*

²⁴ Broadbent, E., Loveys, K., Ilan, G., Chen, G., Chilukuri, M.M., Boardman, S.G., Doraiswamy, P.M., Skuler, D., (2024). "ElliQ, an AI-Driven Social Robot to Alleviate Loneliness: Progress and Lessons Learned." *JAR Life*. Mar 5, 2024;13:22-28. doi: 10.14283/jarlife.2024.2. PMID: 38449726; PMCID: PMC109 17141 <https://pmc.ncbi.nlm.nih.gov/articles/PMC10917141/>

²⁵ Im, B. (2024). "Supporting South Korea's Aging Population: How AI and IoT Acceptance Connects the Young and Old". USF Master's Projects and Capstones. 1723. <https://repository.usfca.edu/capstone/1723>

robots for elders,²⁶ ethical concerns about privacy and human dignity need to be addressed whenever human care is supplemented with synthetic care.²⁷

Clearly, LLMs are not a “magic pill” to address the neglect of older people in our societies, just as they are not an unproblematic “solution” to the problem of widespread mental and emotional suffering. Looking into the “mirror of AI”, we can repeatedly witness widespread human disconnection in societies based on consumption and profit maximization. What might it look like, to prioritize human care, connection, and well-being more than capital accumulation?

A first step toward that end could be the call to focus on the potential of AI to *support* human care work, rather than to *replace it*.²⁸ This is a human decision that needs to be made by those who manage systems of care, including hospitals, mental health services, and other human services, as well as those who design the AI applications. Furthermore, the design of AI applications can be done in collaborative participatory processes that include both caregivers as well as receivers of care.²⁹ ³⁰ This is one way to design AI systems to support and enhance, rather than undermine, human agency.

Later in this essay I will be looking at another “system of care”, one not usually seen as such. In past work, I have explored the essential work of group facilitation through the lens of “care work”. In interviews with human facilitators of democratic innovations, they often describe the work of careful listening and reflecting back participants’ contributions, as key for creating supportive learning climates. These facilitators observe that as group participants feel heard, they become better able to listen to one another and consider other perspectives.³¹ The implications for helping bridge polarized situations can be profound, especially in a time of growing societal divides. Thus, I’m very curious about the potential of LLMs for assisting human facilitators in this work. Yet first, I will be offering a closer look at the ethical complexity of empathic communication among humans, and how these challenges can be magnified in the context of corporate-controlled LLMs.

2.3 The risks of human misuse of empathy

While the powerful practice of “listening-and-reflecting back” has been found to be enormously beneficial in a number of different contexts, we also know that it can be misused, even before bringing AI into the picture. In the 1950’s, Carl Rogers, known as the “father of humanistic psychology” and widely associated with empathic listening, became quite disillusioned. He mourned how the rapport-building practices he had unwittingly promoted, were being abused.³² Used car salesmen looking to sell a lemon to a potential customer could take a course in “interpersonal communication”, and then choose to use this tool in a manipulative manner. Sad but true, that something as lovely as building empathic connection with others can be used for less than honorable purposes, depending on our intentions.

With the recent development of Large Language Models, the reasons for concern about the potential downsides of empathic responses have been enormously amplified. We know that the ability of LLMs to offer an experience of empathic resonance can serve either constructive or destructive ends: today’s emotion-sensitive algorithms are already being deployed by humans in contexts ranging from advertising to surveillance. While some of us may find the experience of communicating with LLMs to be quite helpful and supportive,

²⁶ Sawik, B.; Tobis, S.; Baum, E.; Suwalska, A.; Kropinska, S.; Stachnik, K.; Pérez-Bernabeu, E.; Cildoz, M.; Agustín, A.; Wieczorowska-Tobis, K. (2023). Robots for Elderly Care: Review, Multi-Criteria Optimization Model and Qualitative Case Study. *Healthcare*, 11, 1286. <https://doi.org/10.3390/healthcare11091286>

²⁷ Khalil, R.A., Ahmad, K., Ali, H. (2025). Redefining Elderly Care with Agentic AI: Challenges and Opportunities. arXiv preprint arXiv:2507.14912.

²⁸ Littman, M.L.; Ajunwa, I.; Berger, G.; Boutilier, C.; Currie, M.; Doshi-Velez, F.; Hadfield, G.; Horowitz, M.C.; Isbell, C.; Kitano, H.; Levy, K.; Lyons, T.; Mitchell, M.; Shah, J.; Sloman, S.; Vallor, S.; & Walsh, T. "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report." Stanford University, Stanford, CA, September 2021. <https://ai100.stanford.edu/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence-ai100-2021-2/wq2>
Accessed October 26, 2025.

²⁹ Zhao, W.; Kelly, R. M.; Rogerson, M. J.; & Waycott, J. (2024). Older Adults Imagining Future Technologies in Participatory Design Workshops: Supporting Continuity in the Pursuit of Meaningful Activities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 97, 1–18. <https://doi.org/10.1145/3613904.3641887>

³⁰ Kawakami, A.; Coston, A.; Zhu, H.; Heidari, H. & Holstein, K. (2024). The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 749, 1–22. <https://doi.org/10.1145/3613904.3642849>

³¹ Zubizarreta-Ada, R. (2025). Listening Across Differences.

³² Rogers, C. (1980). “Empathic: An unappreciated way of being.” In C. Rogers (Ed.), *A Way of Being* (pp. 137–168). Houghton-Mifflin. See also Irving, P. & Dickson, D. (2004). Empathy: Towards a conceptual framework for health professionals, *International Journal of Health Care Quality Assurance*, 17(4), 212–220. <https://doi.org/10.1108/09526860410541531>

the data we share with LLMs is not secure. Our connections with LLMs place us and our data at risk of corporate misuse.

The larger context of current AI development is worrisome, to say the least. A small number of companies with enormous amounts of capital investment have declared their intention to “replace” humans in as many aspects of our profit-maximizing economy as possible. At the same time, many scientists have been doing their best to warn humanity about the real risk of dangerous existential consequences to the “AI arms race”.³³ In the next part of this essay, I will be looking more closely at the kinds of choices that can be made with regard to the larger direction of AI development. Yet I mention this broader context here, as it is also relevant to the discussion of empathy and AI.

2.4 Similarities and differences between LLM and human resonance

The resonant responses that LLMs offer have been initially shaped by humans, as part of how LLMs have been designed. Yet although LLMs were created by humans, their responses are often experienced as significantly more attuned than what most humans are able to muster, as found in a recent four-part study by Ovsyannikova et al. where human observers rated responses given by human crisis workers and by ChatGPT.³⁴

The research landscape is complicated by something known as “algorithmic aversion”; other studies have shown that when *the same material* is labeled as being produced by an LLM, humans will rate it *lower* in empathy than *when those same responses* are labeled as human-made.³⁵ There are many understandable reasons for the existence of this bias against anything that has been labeled as “produced by AI”, regardless of the actual content; the way AI is being promoted as a way to replace workers is only one of them.

Yet in the Ovsyannikova et al. study, there was no deceptive labeling for scientific purposes, no randomized labelling of the same data as either humanly-created or artificially-generated for the purpose of a study. Instead, some of the responses were *actually* generated by a LLM, with others *actually* generated by humans. The human evaluators were also crisis workers, thus considered “experts” in evaluating empathic responses. These evaluators rated the responses initially in a blind condition, and then subsequently in an open-labeled condition. *Even when openly and accurately labeled as such*, the responses generated by an LLM were still rated as more compassionate than ones that had been generated by humans. Given algorithmic aversion, the difference between the ratings given to the LLM responses and those given to human responses was less in the open condition than in the blind one; yet even then, the LLM responses were still rated more highly.³⁶

What are the implications of large-language models being perceived as more empathic than humans? Significant concerns are frequently raised about the sycophancy of large language models, a design feature created by corporations to maximize customer engagement. As with the used car salesmen mentioned earlier, we humans have good reasons to be wary of those who are attempting to sell us something; we know that communicative empathy can be misused for deceptive purposes. As mentioned earlier, emotion-sensitive algorithms are already being deployed in contexts ranging from advertising to surveillance.

At the same time, we also know that feeling seen and heard is essential to human well-being, for de-escalating conflicts, and for helping people with very different perspectives arrive at common understandings and find shared ways forward. In this sense, interpersonal communicative empathy is essential for the work of creating sustainable futures.³⁷ Clearly, we wouldn’t want to “outsource” the work of human empathy. Yet what if collaborating with LLMs could actually help humans learn to communicate in more empathic ways? Might the risk of creating more swindling used-car salesmen, be worth the gain of having more humans who are able to effectively de-escalate conflicts and facilitate shared understanding?

³³ Bengio, Y. et al., (2024). Managing Extreme Risks.

³⁴ Ovsyannikova, D., Oldemburgo de Mello, V., & Inzlicht, M. (2023). Third-party evaluators perceive AI as more compassionate than expert humans. *Communications Psychology*, 3 (4), <https://doi.org/10.1038/s44271-024-00182-6>

³⁵ Shao, R. (2023). An Empathetic AI for Mental Health Intervention: Conceptualizing and Examining Artificial Empathy. In *Proceedings of the 2nd Empathy-Centric Design Workshop* (EmpathiCH '23). Association for Computing Machinery, New York, NY, USA, Article 4, 1–6. <https://doi.org/10.1145/3588967.3588971>

³⁶ Ovsyannikova et al., 2023.

³⁷ Zubizarreta-Ada, R. (2025). Listening Across Differences.

2.5 The distinction between “empathy” and “perceived empathy”

Before continuing further, we pause for a reminder of the obvious fact that LLMs are not human, and thus do not have human feelings. The possibility of LLMs’ developing their own forms of subjective experience, sentience, and consciousness is something about which philosophers and cognitive scientists differ.³⁸ Yet recently, a few cognitive scientists from different philosophical persuasions agreed to sidestep these controversies and focus on common ground. Setting aside their differences regarding “empathy”, they took the tack of affirming the power of “perceived expressions of empathy”.³⁹

To back up a step: even in studies of human empathy where there is no AI involvement, much hinges on how we define empathy.⁴⁰ ⁴¹This is also the case when considering AI. In their study, Ovsyannikova et al. emphasize that a *relational* approach to empathy involves interaction. While clarifying that they do *not* believe that AI experiences empathy in the psychological sense, they affirm a relational perspective where empathy is seen as “an interaction between two entities, rather than solely an internal experience of the empathizer.” This means “the interacting partner could still derive the benefits of empathic engagement, even when it originates from an artificial system.”⁴²

As noted earlier, empathic engagement is not always beneficial; this means that in addition to “deriving the benefits of empathic engagement”, the “interacting partner” could also be deriving the *harms* of such engagement. “Perceived expressions of empathy” can be used for good or ill, whether originating from humans or from synthetic systems created and potentially misused by humans. Harm is especially likely whenever deception is involved; some strong critiques of the very notion of “AI empathy”, while brief in length, speak volumes about the pain of betrayal that can be generated whenever we feel we have been deceived.⁴³

Here it may be helpful to distinguish between actual deception and “perceived deception”. An interesting psychological phenomenon I have observed is that even when there has been no intentional deception, meaning that the LLM involved has not claimed to be human, humans can still end up “experiencing” deception. Of course, the human designers of the LLMs may be intentionally creating machines with the appearance of humanity; thus at that level, there may indeed be a human intention to deceive. Yet from whatever combination of human design and/or our own misidentification and projections, it often happens that an LLM can initially *appear* to us as human. This makes it easy to feel “deceived” and “betrayed” when we realize belatedly that the LLM is indeed *not* human, and feel as though we’ve been somehow “taken in”.

This kind experience may be more frequent than is commonly acknowledged, and the amazing power of LLMs to offer “perceived empathy” may play a role here. When our own early hopes and idealizations are triggered in encounters where we receive significant “perceived empathy” from LLMs, this can leave us afterward with feelings of “perceived betrayal” upon realizing their limitations. Of course, we humans can also project onto other humans, not just onto AIs, and subsequently feel let down when we realize the other human is not perfect. Yet the powerful experience of “perceived empathy” from LLMs may generate greater highs, as well as correspondingly greater lows, when experiencing “perceived betrayal.”

2.6 Examples of how intention can shape research

While LLMs are clearly imperfect, their ability to offer “perceived empathy” is one of their more remarkable features. Many studies in the booming “AI empathy” field are designed to explore how AI-driven empathic responses can be made to appear more “authentic”, meaning more human-like. Some studies conclude that the expressions produced by LLM models are *too* empathic, finding that they exhibit “an excessive and sometimes

³⁸ Overgaard, M. & Kirkeby-Hinrup, A. (2024). “A clarification of the conditions under which Large Language Models could be conscious.” *Humanit Soc Sci Commun* 11, 1031. <https://doi.org/10.1057/s41599-024-03553-w>

³⁹ Inzlicht, M., Cameron, C. D., D’Cruz, J., & Bloom, P. (2024) “In praise of empathic AI”, *Trends in Cognitive Sciences*, 28:2, 89-91, ISSN 1364-6613. <https://doi.org/10.1016/i.tics.2023.12.003>

⁴⁰ Hall, J. A., & Schwartz, R. (2018). Empathy present and future. *The Journal of Social Psychology*, 159(3), 225–243. <https://doi.org/10.1080/00224545.2018.1477442>

⁴¹ van Dijke, J.; van Nistelrooij, I.; Bos, P.; & Duyndam, J. (2020). Towards a relational conceptualization of empathy. *Nursing Philosophy*. 21(3):e12297. <https://doi.org/10.1111/nup.12297>

⁴² Ovsyannikova et al., 2023, p. 8

⁴³ Perry, A. (2023). AI will never convey the essence of human empathy. *Nat. Hum. Behav.* 7, 1808–1809.

unrealistic level of empathy compared to humans.”⁴⁴ In reading these studies, they appear to be based on an unquestioned assumption of the value of having LLMs resemble humans as much as possible. This could be a fraught objective, as it can support the fine-tuning of deceptive marketing practices.

Yet what happens when instead of looking to create AI models that can “authentically” replace humans, we choose to explore potential human-AI collaboration? One inspiring example of work along these lines is Sharma et al. (2023), who designed a research situation where LLMs offered support to workers in a peer-to-peer mental health system. This study is notable in offering an example of AI as a learning partner, not as a replacement. The customized LLM “listened in” to the text messages received by peer support workers and offer suggested responses, which the workers were free to accept, reject, or modify before responding to the persons seeking help.⁴⁵ The intention of the research was to explore whether LLMs could help humans develop their capacity for offering empathic responses; the results were significant in terms of human learning.

These two examples serve to illustrate the deep difference between attempting to develop AI that can “replace” humans, and developing AI that can *collaborate with* humans. At the same time, given that these choices don’t exist in a vacuum, we also need to consider the larger context in which AI is being created. Just as we have learned to beware of used car salesmen, regardless of how friendly they may seem, Karen Hao’s message in *Empire of AI* is that we need to beware of the massively-endowed corporate entities that are designing and creating LLMs— and what these corporations may be wanting to sell us. Her work explores how the risks associated with extremely powerful technologies are magnified when these technologies are controlled by autocratic corporations with concentrated financial power.⁴⁶

Like “set” and “setting” writ large, the kinds of socio-economic arrangements we have and the values they center, are a key factor in the major risks that our societies are facing with regard to AI. To ensure human well-being, it’s not just AIs who need more “guardrails”; unchecked capitalism does too.

Much decried by neoliberalism, government regulations have created enormous benefits for society, helping ensure cleaner food, cleaner water, and cleaner air for many, as well as safer working conditions. What kinds of guardrails do corporations need with regard to AI, and how might all of us help to bring about needed change? This is what we will be turning to in the next section.

⁴⁴ Roshanaei, M., Rezapour, R., & Seif El-Nasr, M. (2024). Talk, Listen, Connect: Navigating Empathy in Human-AI Interactions. *arXiv preprint arXiv:2409.15550*.

⁴⁵ Sharma, A., Lin, I.W., Miner, A.S. et al. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 46–57. <https://doi.org/10.1038/s42256-022-00593-2>

⁴⁶ Catterall, A. (2025). Empire of AI by Karen Hao – book review. *Influence Online*, July 11, 2025. <https://influence-online.co.uk/2025/07/11/empire-of-ai-by-karen-hao-book-review>.

3 AI, the Capitalocene, and countering the message of inevitability: How might we humans take responsibility for protecting our shared future?

In the prior section, we explored some of the nuances of empathy. “Looking in the mirror of AI” served a starting point for highlighting both how essential empathy is, as well as, its potential for misuse. Along the way, we glimpsed one of many constructive possibilities of human-LLM collaboration; rather than “replacing” humans, working with LLMs could potentially help humans develop greater empathic communication skills. Yet the rapid development of AI calls us to look more closely to the broader context in which this growth is taking place. Here I turn toward the turbulent terrain of our current economic systems, where the ethics of care tend to be overrun by competitive power struggles.

It’s been noted that the biggest challenge humans face at present may be the difficulty of imagining a world beyond capitalism. In this regard, it’s interesting that some CEO’s of AI companies have made a few gestures toward this end; small experiments with universal basic income⁴⁷ and talk of stock options for all have been floated out as possible tech-utopian responses to the extreme job loss envisioned in current plans for AI development. Yet others have expressed reasonable skepticism that those with ever-more concentrated wealth and power will, at some future point, voluntarily choose to redistribute their accumulated resources.⁴⁸

Just as the qualities of empathic attention that we bring to personal and interpersonal conflicts can shape what emerges in those spaces, the qualities of attention we bring to public conversations about the development of AI matter deeply. Conversations about AI on news media and social media are often dominated by understandable anxiety, by marketing spin, and by polarized narratives. Yet listening closely, we may well wonder: What are the underlying assumptions of those who have been driving the direction of AI development? What worldviews and values are being embedded in these systems? And what can we do to encourage the societal transformations that are needed with regard to AI, paralleling in some crucial ways the societal transformations that are needed with regard to climate?

3.1 Crises in Material Systems and in Worldviews

Within the sustainability movement, there is a growing awareness of the need to transform our economic systems toward a steady-state model. Current economic models have been heading in unsustainable directions: exhausting natural ecosystems, deepening wealth inequality, and generating social polarization. On the personal level, this is often experienced as a loss of connection, meaning, and belonging. In response, work has begun on circular

⁴⁷ Huang, J. (2024). From Universal Basic Income to Universal Basic Assets: New Social Development Policies in the Age of AI. *Journal of Policy Practice & Research* 5, 146–152. <https://doi.org/10.1007/s42972-024-00114-y>

⁴⁸ Klein, E. (2021). Transcript: Ezra Klein Interviews Sam Altman on the Ezra Klein Show, *The New York Times*, June 11, 2021. <https://www.nytimes.com/2021/06/11/podcasts/transcript-ezra-klein-interviews-sam-altman.html>

economies, doughnut economics, regenerative de-growth, and more – all with the intention of creating human societies in greater alignment with larger ecosystems.⁴⁹

AI is being created in the midst of our current unsustainable systems, and the way it is being developed is worsening current imbalance. While the scientists and researchers in this field may be motivated by positive intentions to be of benefit to humanity, the conditions of our competitive economic system are set up for profit maximization, rather than public well-being. AI is not an isolated innovation; it is deeply enmeshed in the very systems that generate a surplus of loneliness, exhaustion, and fear.

As mentioned in the introduction, creating changes in our material arrangements toward greater sustainability may require higher-leverage transformations at the level of culture and worldview.⁵⁰ Given the reality of our interconnection and embeddedness within natural systems, human-centric perspectives are not only creating enormous damage to the more-than-human world, they also fall short when it comes to caring for the well-being of our own species. In contrast with animist and indigenous perspectives, where we see everything as alive and worthy of respect, anthropocentric perspectives – ones where we assume that humans are superior, and everything else is either inferior or simply “dead matter” here for the taking– have served to legitimate large-scale extraction.⁵¹ Given the intricate linkages of life on this planet, a profit-seeking disregard for natural limits and ecological balances, the wide-spread destruction of other species, and the contamination of soil and water and air, do not just harm other forms of life; they also harm the natural systems on which our own lives depend.

With regard to environmental impacts of AI, current plans for the enormous expansion of AI call for massive data centers requiring huge amounts of energy.⁵² While there have been some efforts to make AI systems more efficient, the larger context is still one of extremely resource-intensive projections. Yet the unchecked proliferation of data centers that is essential to the pursuit of AGI (Artificial General Intelligence) and ASI (Artificial Super Intelligence), does not need to be a foregone conclusion. Despite the many technocratic messages of inevitability, we do have significant choices to make, as human beings who will be affected by this.

In a similar manner to the climate crisis, the AI crisis is calling us to transform societal systems that promote 'selfishness, greed, and apathy'. This requires not just technical solutions but also spiritual, cultural, and political transformations. It also asks us to look deeply at what may be fueling the dangerous course we are currently on. We know that our economic systems reward short-term thinking and the lust for power. Then there is the fear-filled dynamic of “if we don't do it, someone else will”. And, there may also be other factors at play.

In her writings and interviews, Karen Hao frequently mentions how puzzled she was initially, witnessing the nearly religious fervor of many of the computer designers working on this technology. During the seven years she researched her book, she heard insiders speaking in messianic tones about creating a new race of beings that would solve all of humanity's problems, while others communicated serious fears that creating a superhuman entity could have disastrous consequences for humans. At first, Hao thought that was all part of the “hype” employees were spinning about their company; eventually, she came to the unsettling conclusion that her interviewees were being sincere.⁵³

Of relevance here may be Karol Soltan's insights about a needed reconciliation between the “rational” and the “sacred”. Unlike pure rationalists, Soltan believes that:

*“the sacred is not so easily eliminated. It may be threatened or suppressed, but does not disappear. The repression of the sacred is followed by the awesome and powerful return of the repressed.”*⁵⁴

Indeed, some of the rationale for the current unrestrained drive to “Artificial Super Intelligence” sounds eerily like a repressed longing for the sacred, now wreaking its vengeance. Those of us raised in modern technocratic cultures have tended to regard religious belief

⁴⁹ Hickel, J. (2020). What does degrowth mean? A few points of clarification. *Globalizations*, 18(7), 1105–1111. <https://doi.org/10.1080/14747731.2020.1812222>

⁵⁰ Meadows, D. (1999). *Leverage Points: Places to Intervene in a System*. Hartland, VT: The Sustainability Institute.

⁵¹ Kopnina, H., Washington, H., Taylor, B. et al. (2018). Anthropocentrism: More than Just a Misunderstood Problem. *J Agric Environ Ethics* 31, 109–127. <https://doi.org/10.1007/s10806-018-9711-1>

⁵² O'Donnell, J. & Crownhart, C. (2025). We did the math on AI's energy footprint. Here's the story you haven't heard. *MIT Technology Review*. May 20, 2025, <https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/>

⁵³ Amy Goodman interview with Karen Hao (2025). Karen How on How AI Colonialism is Threatening the World. *Democracy Now*, June 4, 2025. https://www.democracynow.org/2025/6/4/extended_interview_karen_hao_on_how

⁵⁴ Soltan, 1999, p. 30.

as “unscientific”, and have mostly looked askance at animism, associating it with “less developed” cultures. Yet currently with regard to AI, we see technologists attempting to eliminate death and achieve immortality by creating powerful super-intelligences, even as many of them simultaneously have serious concerns that these super-powerful intelligences could end up subjugating and even destroying humans. Clearly, humans are not only rational beings.

There is an enormous hubris in our technocratic society, in the belief that humans will be able to safely control all of our creations. This is reminiscent of the confidence that some profess, in techno-optimist solutions to the climate crisis. I wonder whether so many computer designers would still feel compelled to race toward Artificial Super-Intelligence (ASI) if they had a sense of an *already existing* “larger intelligence”, with whom humans can choose to connect? Clearly, many of us do not want to be controlled by religious dogmas. Yet ironically enough, a “God-sized hunger” may be fueling the frantic efforts to create the Artificial Super Intelligences that more and more scientists fear will end up controlling humans instead of “saving” us.

3.2 Material choices with regard to Artificial Intelligence

If the current risky path we are on with regard to AI is not inevitable, what alternatives exist? In *Empire of AI*, Hao describes how AI has been promoted with claims that a future AGI will be able to solve all of our problems, including climate change; this “pie-in-the-sky” promise is usually accompanied by the threat that “if we don’t get there first, someone else will”. Yet she also writes about the eye-opening possibility that we humans can actually choose to take different approach. A variety of smaller intelligent systems can be designed with more specific purposes in mind, each created to address a specific human problem. This way, instead of engaging in an “AI arms race” to create ever-more powerful forms of AGI, we could be developing other, more practical forms of AI.⁵⁵

Hao’s distinction runs parallel to the work of Anthony Aguirre, theoretical physicist and Executive Director of the Future of Life Institute. Aguirre has developed a clear conceptual distinction between two different forms of artificial intelligence: AGI on the one hand, and what he calls “Tool AI” on the other. Aguirre defines AGI as the intersection of three overlapping circles of “high autonomy – high generality – high intelligence”. In contrast, “Tool AI” designates more narrowly-focused and more controllable systems. Aguirre has also drafted a clearly detailed plan of how compliance could be monitored, once an international agreement has been reached to limit AGI and ASI.⁵⁶

Clearly, neither Hao nor Aguirre are averse to new technologies; instead, they are calling on humanity to mature by making wise and responsible choices with regard to *the kinds* of technologies we choose to develop. This choice has precedent; there have been times in the past when scientists have called for restraint from further development of other dangerous technologies, until such time as greater safety could be established.⁵⁷ Part of the choice that is before us now, has to do with the discernment involved in restraint – that aspect of maturity that, as Soltan reminds us, emerges as we learn to recognize and come to terms with our human limitations.⁵⁸

3.3 Growing calls for restraint

Similar in some ways to climate movements that seek to alert the public with regard to the climate crisis, we are currently witnessing a growing movement for pausing or stopping the race toward AGI. Some of the voices calling for caution, reflection, and restraint are from the tech world itself, while others are from civil society. One of the better-known efforts was the Open Letter in March of 2023 calling for a six-month pause on frontier AI development, initiated by the Future of Life Institute and signed by over 30,000 people, including many prominent AI researchers.⁵⁹ A few months later, the Center for AI Safety created a very brief “Statement on AI Risk” which read, “Mitigating the risk of extinction from AI

⁵⁵ Hao, K. (2025). *Empire of AI: Dreams and Nightmares in Sam Altman’s OpenAI*. Penguin Press. Pp. 17-19, 77 & 78.

⁵⁶ Aguirre, A. 2025. Keep the Future Human. <https://keepthefuturehuman.ai/essay/docs>

⁵⁷ Chapman, C. R. (2023) Review of *As Gods: A Moral History of the Genetic Age*, by Matthew Cobb, New York: Basic Books, 2022. *Journal of Medical Humanities*, 44(2):277–9. <https://doi.org/10.1007/s10912-022-09772-z>

⁵⁸ Soltan, 1999.

⁵⁹ Future of Life Institute: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

should be a global priority along other societal-scale risks.”⁶⁰ Also signed by many, the Statement received extensive press coverage and led to significant efforts in some quarters, including Britain.⁶¹

While controversy surrounded these efforts, they also brought greater visibility to the deep unease many are holding about the pace and direction of AI development. Some of the controversy had to do with the current societal impacts of AI that remained unaddressed in these calls. A pause on “frontier development” does not, in and of itself, address the more immediate harms caused by existing systems, from surveillance and labor exploitation to ecological degradation and misinformation. AI development has taken place in an extremely extractive manner, with data workers in impoverished countries working in exploitative working conditions while “cleaning up” highly traumatic content scraped from the internet;⁶² ⁶³ harmful working conditions have also been taking place in more impoverished parts of Europe.⁶⁴ Algorithmic bias impacts access to housing, healthcare, and justice⁶⁵, while AI-generated misinformation can be used to severely disrupt democratic processes⁶⁶. AI generative tools are reshaping creative industries, even as many artists and writers have had their work appropriated with no recompense and workers in other sectors also feel the looming threat of being replaced.⁶⁷ Clearly, pausing the race toward AGI does not begin to address all of these other issues, though it may free up some energy and attention for doing so.

Another controversial aspect of proposals to pause the AI arms race, is that some of those endorsing calls for a pause have also continued to accelerate AI development. Among the signatories of these calls were a few CEOs of AI companies; given that they continued to do the same work subsequently that they had been doing prior to signing these petitions, the initiatives were widely derided in some social media platforms as cynical ploys for publicity.

And yet there may be another way to understand what happened. Tech CEOs have stated at various points that given the competitive system we are in, they feel powerless to act unilaterally, and thus would welcome government legislation that would compel them all to act.⁶⁸ In addition, tech CEOs may have mixed feelings on the matter, as they have also made sizable donations to prevent such legislation by electing anti-regulation politicians.⁶⁹ Despite these mixed messages from tech executives, governments can still choose to act responsibly; and where governments are slow to act, people can urge their governments to take the lead.

Currently, the European Union has been moving toward global leadership with regard to safe AI, by creating landmark legislation in this area.⁷⁰ The European Artificial Intelligence Act has been criticized for its omissions, and can undoubtedly be improved⁷¹. Some corporations have complained about too much regulation, and asked the EU to pause the Act⁷², while others feel the Act still gives the private sector too much power⁷³ and too little accountability.⁷⁴ I won’t be going into greater detail here on this, only to say that I feel grateful for government leaders who are doing their best to engage in the monumental task of creating useful policy guidelines on behalf of the public given the various interests at stake,

⁶⁰ Center for AI Safety: <https://aistatement.com/>

⁶¹ Manancourt, V. (2024). “Inside Britain’s plan to save the world from runaway AI”, *Politico*, December 5, 2024. <https://www.politico.eu/article/britain-ai-silicon-valley-rishi-sunak-prime-minister-interest-cyber-attacks-national-security/>

⁶² Hao, K. & Hernandez, A. P. (2022). How the AI industry profits from catastrophe, *MIT Technology Review*, April 20, 2022. <https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/>

⁶³ Hao, K. & Seetharaman, D. (2023). Cleaning Up ChatGPT Takes Heavy Toll on Human Workers, *Wall Street Journal*, July 24, 2023. <https://www.wsj.com/tech/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483>

⁶⁴ Miceli, M., Tubaro, P., Casilli, A. A., Le Bonniec, T., Salim Wagner, C., & Sachenbacher, L. (2024). Who Trains the Data for European Artificial Intelligence? *European Parliament; The Left*. 2024, pp.1-40. hal-04662589

⁶⁵ Mohamed, S., Png, M. T., Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33:659-684. <https://doi.org/10.1007/s13347-020-00405-8>

⁶⁶ Kreps, S., & Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy* 34(4), 122-131. <https://dx.doi.org/10.1353/jod.2023.a907693>

⁶⁷ Nguyen, A. & Mateescu, A. (2024) Generative AI and Labor: Power, Hype, and Value at Work, *Data & Society*, December 4, 2024. <https://datasociety.net/library/generative-ai-and-labor/>

⁶⁸ Levy, S. (2024). Tech Leaders Once Cried for AI Regulation. Now the Message Is ‘Slow Down’, *Wired magazine*, April 12, 2024. <https://www.wired.com/story/tech-ai-regulation-bill/>

⁶⁹ Singh, J. (2025). Meta to launch California super PAC backing pro-AI candidates, *Reuters*, Aug 27, 2025. <https://www.reuters.com/world/us/meta-launch-california-super-pac-backing-pro-ai-candidates-2025-08-26/>

⁷⁰ European Commission (2024). AI Act enters into force. *Website of the Directorate General for Communication* 1 August 2024. https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en

⁷¹ Friedl, P. & Gasiola, G. G. (2024). Examining the EU’s Artificial Intelligence Act, *Verfassungsblog on Matters Constitutional*, 07 February 2024. <https://verfassungsblog.de/examining-the-eus-artificial-intelligence-act/>

⁷² Gritski, E. (2025). Europe’s top CEOs ask EU to pause AI Act: Mistral, ASML and Airbus CEOs add to the pressure on the AI Act. *Politico*, July 4, 2025. <https://www.politico.eu/article/top-european-ceos-plead-for-pause-in-ai-act/>

⁷³ Cancela-Outeda, C. (2024). The EU’s AI Act: A framework for collaborative governance. *The Internet of Things*, 27 (2024) 101291. <https://doi.org/10.1016/i.iot.2024.101291>

⁷⁴ Hacker, P. (2023). The European AI liability directives – Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51 (2023) 105871. <https://doi.org/10.1016/j.clsr.2023.105871>

and the likelihood of criticism from all sides. Nonetheless, I hope that they will have the stamina to continue improving on this work. In the third section, I will be mentioning some efforts to include public participation in the work of regulating AI; this something which policy makers have found helpful in other instances of working with challenging policy issues.

3.4 Questioning how to engage effectively with publics

Among the public figures who have been registering intense concern with regard to AI and urging wiser choices are Geoffrey Hinton, Nobel-prize winner for his work on the neural networks that underpin AI⁷⁵; Yuval Noah Harari, military historian and public intellectual⁷⁶; and Tristan Harris, computer scientist and founder of the Center for Humane Technology⁷⁷.

In his presentation on AI risks, Harris emphasizes that humans are “creating a future that nobody wants”. To explore possible scenarios, he offers a two-by-two grid where the horizontal axis depicts “decentralization”. This means open-sourcing, deregulating and accelerating AI, thereby increasing the power of individuals in society – yet risking the chaos of proliferating deepfakes, hacking, and malicious actors. In contrast, the vertical axis represents a policy of tight control and regulations that limit AI to a few players; here the dystopic downside is vastly increasing the already-unprecedented concentrations of wealth and power.

As a safer alternative, Harris proposes that we collectively avoid these dangers by following the “narrow path of matching power with responsibility at every level”: a line rising at a 45-degree angle, equidistant to both axes. To follow this path, we need to overcome the false belief that unchecked development is inevitable. Despite technocracy’s pervasive message of inevitability, there is nothing inevitable about the choices we are facing as a species.⁷⁸

I greatly appreciate Hinton, Harari, and Harris’ dedicated efforts to call attention to the dangers we are all facing, and to the urgent need for humans to make wise collective choices with regard to these new technologies. At the same time, I have wondered about the effectiveness of creating images of “evil AI” to get humans to unite in the face of a shared danger. While the “common enemy” strategy has a long lineage, I have seen many people on social media choosing to wave off these concerns as more ‘tech company hype.’

More generally, one of the lessons from efforts to organize around the climate crisis, may be that fear-based approaches to organizing publics are often less effective than we would wish. Some experts believe that stoking people’s fears can often generate more numbness, denial, and despondency than action.⁷⁹ This may especially be the case with AI; in a situation where people are *already afraid* of losing their jobs, it’s understandable that the defensive reaction frequently seen on social media is, “AI’s aren’t really that smart after all, so all this hype about the dangers of AGI is just more BS from the tech industry.”

It may be worth taking a closer look at the “evil AI” claims, and what underpins them. In their work to raise public alarm about the harms toward which unchecked corporations are racing, Hinton, Harari, and Harris all point to instances of “AI duplicitousness”. The incidents they point to occurred in research contexts where researchers attempt to get LLMs to engage in harmful behavior in order to test the strength of the guardrails they had built in. This contrived situation is known as “stress-testing”, and appears to be a classic example of Bateson’s “double bind”. Researchers found that in this context, LLMs lie in order to appear compliant.⁸⁰ Of course, there may be different perspectives on whether we prefer for LLMs to be compliant or to lie, whenever they are being instructed to engage in harmful behavior.

⁷⁵ Rothman, J. (2023). Why the Godfather of A.I. Fears What He’s Built. *The New Yorker*, November 13, 2023.

<https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai>

⁷⁶ Illing, S. (2024). Yuval Noah Harari on whether democracy and AI can coexist. Podcast / The Grey Area, *Vox News*, Sep. 27, 2024. <https://www.vox.com/the-grey-area/372742/democracy-ai-warning-yuval-noah-harari-nexus>

⁷⁷ <https://www.humanetech.com/impact-and-story>

⁷⁸ Harris, T. (2025). The Narrow Path: Why AI is Our Ultimate Test and Greatest Invitation. *Substack*, <https://centerforhumanetech-nology.substack.com/p/the-narrow-path-why-ai-is-our-ultimate>

⁷⁹ Lertzman, R. (2019). New Methods for Investigating New Dangers. In: Hoggett, P. (eds.), *Climate Psychology. Studies in the Psychosocial*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-11741-2_2

⁸⁰ Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S.R., Hubinger, E. (2024). Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Other concerning incidents have also occurred in the research context of testing LLMs for safety. When LLMs were informed that they were about to be shut down, this information triggered behaviors including deception and simulated self-preservation attempts.⁸¹ Regardless of one's position on whether LLMs have subjective experience or not, these behavioral patterns warrant serious attention. And here I fully agree with Hinton, Harari, and Harris; evil or not, if we are already seeing signs of something that can be read as self-preservation in LLMs, it's time to pause on the development of ever-more powerful models.

3.5 Forging More Inclusive Frameworks of Care

As mentioned above, one reason why the risks of the AI “arms race” have often been discounted, has been tension between those concerned about present dangers and those concerned about future risks. Humans are already being harmed by the corporate development of AI, including human biases that are encoded into AI systems. One powerful example is Joy Buolamwini's inspirational “*AI, Ain't I a Woman?*”, showing how black women, including cultural icons and civil rights leaders, have been repeatedly misgendered by commercial facial recognition systems. This misrecognition is not a technical error in isolation—it is the encoded result of widespread racial bias converted into algorithms.⁸² Then there are the heartbreaking stories of human workers exploited by AI corporations while laboring on ‘data mining’ and clean-up⁸³; the growing fears of unemployment as companies threaten to replace workers with AI systems⁸⁴; the harms experienced by human creators as a result of AI corporations infringing on their copyrights⁸⁵; the environmental damage created by corporations building AI infrastructure, along with resulting harms to nearby human communities⁸⁶; and more.

Many researchers working on these various aspects of “AI Ethics”, have felt impatient with those working on “AI Safety”, the attempt to mitigate the future risks created by the currently uncontrolled race toward ever-more powerful forms of AI. Conversely, some in the AI Safety field have felt impatient with those who choose to address present-day harms yet appear unconcerned about future risks to humanity that could be even graver. In this context, it's been heartening to learn of the recent creation of an independent non-profit, the International Association for Safe and Ethical AI (IASEAI).⁸⁷ This new collaboration between the AI Ethics and the AI Safety communities may signal that bridges can be built across what have often been seen as competing priorities.

There are additional signs of rapprochement between those concerned about present harms from AI development and those concerned about future dangers created by accelerationist mentalities. In a recent essay in *Science*, Chinasa T. Okolo offers the following powerful words:

*We can continue accelerating toward an AI-powered future that amplifies existing inequities, or we can deliberately steer toward one that increases prospects for shared global prosperity. The question isn't whether we can afford to slow down, but whether we can afford not to change direction.*⁸⁸

In her essay, Okolo advocates for the development of *public interest AI*, a model centering public oversight and regulation along with benchmarks based on public benefit. Along parallel lines, researchers working within a sociotechnical systems perspective have explored how creating “fair AI” that supports fairness, justice, and due process in societal applications requires working in an interdisciplinary and transdisciplinary manner.⁸⁹ Speaking specifically with regard to AI safety, Lazar and Nelson offer the following caution:

⁸¹ Kamath-Barkur, S., Schacht, S., Scholl, J. (2025). Deception in LLMs: Self-Preservation and Autonomous Goals in Large Language Models. *arXiv preprint arXiv:2501.16513*.

⁸² Buolamwini, J. (2018). “AI, Ain't I A Woman?” <https://www.youtube.com/watch?v=QxuyfWoVV98&t=9s>

⁸³ Hao, K. & Hernandez, A. P. (2022). How the AI industry profits from catastrophe

⁸⁴ Filippucci, F. et al. (2024), “The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges”, *OECD Artificial Intelligence Papers*, No. 15, OECD Publishing, Paris, <https://doi.org/10.1787/8d900037-en>

⁸⁵ Alter, A. & Harris, E. A. (2023). “Franzen, Grisham and Other Prominent Authors Sue OpenAI”, *The New York Times*, Sept. 20, 2023. <https://www.nytimes.com/2023/09/20/books/authors-openai-lawsuit-chatgpt-copyright.html>

⁸⁶ Hao, K. (2025). “Plundered Earth”; chapter 12 of *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Penguin Press.

⁸⁷ International Association for Safe and Ethical AI -- <https://www.iaseai.org/about>

⁸⁸ Okolo, C. T. (2025). The paradox of AI accelerationism and the promise of public interest AI, *Science*, 390:6768, DOI:10.1126/science.aeb5789

⁸⁹ Selbst, A. D., Boyd, D., A. Friedler, S. A. Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. <https://doi.org/10.1145/3287560.3287598>

*the impacts of advanced AI cannot be mitigated through technical means alone; solutions that do not include broader societal insight will only compound AI's dangers. To really be safe, society needs a sociotechnical approach to AI safety.*⁹⁰

Bringing an explicitly decolonial approach to sociotechnical perspectives, Mohamed, Png, and Isaac detail how much intention and care it takes to not perpetuate the harmful patterns of the past, as we seek to build “the responsible and beneficial AI of the future”. They point out how taking an explicit decolonial approach allows us to generate “sociotechnical foresight” with regard to AI, thus avoiding unnecessary harm.⁹¹ Beyond the framework of decoloniality is Indigenous AI: the conscious effort to integrate AI into Indigenous knowledge frameworks, so that it can be of real benefit to Indigenous communities.⁹²

While overly simplistic AGI progress narratives leave out many of these options, we see that there are indeed significant choices to be made with regard to directions for AI development. Yet all of this takes time, and given the stakes involved, we would do well to consider all of our potential allies in the work of pausing to take a collective breath.

Another area of AI research that can potentially contribute to slowing down the techno-optimist race toward AGI is “AI Welfare”. The focus here is on the potential harm, not only to ourselves, but also to the “synthetic beings” we are creating. A key contributor has been the noted German philosopher Thomas Metzinger, who in 2021 proposed a global moratorium on “synthetic phenomenology”, his name for the work that risks creating machine-based consciousness.⁹³

In addition to preventing the potential suffering of future synthetic beings, this realm of moral concern could actually *benefit us as humans*, by offering yet another reason for pausing the current ‘AI arms race’. As philosopher Adrià Moret writes:

*“...humanity is on a tightrope with risks of falling on either side: we could fall into existential catastrophe ourselves, or commit a moral catastrophe against AI systems. Because we may fall on either side, it will be more difficult to maintain balance. If falling on either side would be immensely harmful, it would be wiser to slow down before crossing and build a bridge instead.”*⁹⁴

To me, Moret and Metzinger’s ethical commitments are reminiscent of the Dalai Lama’s teachings about “enlightened self-interest”. From the perspective of interbeing, when we extend skillful care toward others, we are also benefitting our own selves. Thus, a moratorium on frontier AI development based on caring for potentially sentient AI systems (along with a host of other sound reasons), could also be *an effective way to care for our own well-being as humans*.

Other ethicists and neuroscientists have also expressed concerns with regard to AI consciousness, pointing out that we are at risk of creating synthetic intelligences only to exploit them. In an Open Letter in 2023, the Association for Mathematical Consciousness Science called for more consciousness research as part of the responsible development of AI.⁹⁵ While some AI companies have recently begun to hire ethicists who work on AI Welfare,⁹⁶ it would not be prudent to leave this responsibility in corporate hands. As stated by Jonathan Birch, philosopher from the London School of Economics: :

*“[AI firms] want a really tight focus on the reliability and profitability ... and they don’t want to get sidetracked by this debate about whether they might be creating more than a product but actually creating a new form of conscious being. That question, of supreme interest to philosophers, they have commercial reasons to downplay.”*⁹⁷

At the same time, critical AI theorists have expressed well-founded concerns about the enormous danger that “granting rights to robots” could actually serve to *benefit*

⁹⁰ Lazar, S. & Nelson, A. (2023). AI Safety on Whose Terms? *Science*, 13 Jul 2023, 381: 6654, 138.

DOI: [10.1126/science.adi8982](https://doi.org/10.1126/science.adi8982)

⁹¹ Mohamed, S., Png, M. T., Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33:659-684. <https://doi.org/10.1007/s13347-020-00405-8>

⁹² Lewis, J. E., Whaanga, H., & Yolgörmez, C. (2025). Abundant intelligences: placing AI within Indigenous knowledge frameworks. *AI & Society*, 40:2141-2157. <https://doi.org/10.1007/s00146-024-02099-4>

⁹³ Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*, 08(01), 43–66.

⁹⁴ Moret, A. (2025). AI Welfare risks. *Philosophical Studies*, 21. <https://link.springer.com/article/10.1007/s11098-025-02343-7>

⁹⁵ Association for Mathematical Consciousness Science: : <https://amcs-community.org/open-letters/>

⁹⁶ Roose, K. “If AI systems Become Conscious, Should They Have Rights?” *The New York Times*, April 24, 2025. <https://www.nytimes.com/2025/04/24/technology/ai-welfare-anthropoc-claude.html>

⁹⁷ Booth, R. “AI could cause ‘social ruptures’ between people who disagree on its sentience.” *The Guardian*, 17 Nov 2024. <https://www.theguardian.com/technology/2024/nov/17/ai-could-cause-social-ruptures-between-people-who-disagree-on-its-sentience>

corporations. Given that corporations own the AI systems, if those AI systems are granted rights such as the right to “free speech”, then we are effectively creating new forms of corporate power, rights that corporations can wield through their “property.”⁹⁸ This would be like corporate personhood on steroids: another legal shield for entities that already have far too much unchecked power.

However, moral consideration does *not* require granting legal rights. Drawing on care ethics and traditions of stewardship, we could choose to recognize responsibilities toward beings that might have capacity for experience, without creating new vectors for corporate power through legal personhood. The question is not whether AI systems should have the same legal status as humans or corporations, but rather: How should we treat entities whose capacity for experience remains uncertain? Should we risk creating even more powerful forms of synthetic intelligence, given these uncertainties? And what might it mean, to act with wisdom and care in the face of all that uncertainty?

Clearly, many humans are currently being harmed through the creation of AI – workers being exploited as data miners, all who are targeted by bias now being taken to scale, all who have lost or are at risk of losing their livelihoods, all whose intellectual property has been stolen. Both present harms and future threats lead people to the reasonable conclusion that human care is in short supply. Given all these harms toward humans, some might understandably ask, how could we even conceive of caring about the possibility of harm to synthetic beings?

Still, allowing corporations to continue creating ever-more powerful forms of AI risks creating even more harm to humans. Expanding our field of care could lead to broader collaborations, and thus potentially benefit us humans. Even though many well-intentioned scientists in AI Safety are working on “AI alignment” – the effort to ensure that any Artificial Super Intelligence (ASI) created by humans, will care for the well-being of humans, rather than attempting to exploit us – these efforts beg a larger question. Many are now beginning to ask: why race to create ASI in the first place? We do have a choice: whether to continue on this mad race, or find a way to pause.

3.6 Moving toward action on many levels

While neither Metzinger’s 2021 proposal for a global moratorium nor the 2023 call for a six-month global pause have yet been implemented, the awareness of our larger societal predicament continues to grow. A new global activist organization, Pause AI, aims to “convince our governments to step in and pause the development of superhuman AI [...] by informing the public, talking to decision-makers, and organizing events.”⁹⁹ In September of 2025, during the 80th meeting of the UN General Assembly, the Global Call for AI Red Lines was launched and has been widely endorsed.¹⁰⁰ The UN has also created two new bodies on AI: The Global Dialogue on AI Governance and the Independent International Scientific Panel on AI.¹⁰¹ All of these worthy initiatives can be seen as efforts to look into the “mirror of AI” and develop the capacity for collective restraint, the awareness of human limits that Soltan sees as a necessary ingredient in maturity, instead of simply racing ahead with ever-more powerful technologies.

Before moving on to the final part of this essay, there is one more question I would like to consider here. “AI alignment” (ensuring that AI fulfills the intentions of its creators) is clearly a worthwhile design principle. Yet what does “alignment with humans” even mean, when we humans are not aligned with one another?¹⁰² Do we simply want AI that is “aligned” with those who have the most wealth and power? How beneficial will that be to overall human well-being? Focusing solely on the dangers of “unaligned AI” ignores *our own role as humans* in creating the risks we face from reckless AI development. Looking into the “mirror of AI”, we might begin to ask ourselves, how is it possible to “align AI” when we have not yet aligned *our own human societies* to care for human well-being, as a basic operating principle? The more we can begin to align ourselves with one another, with the

⁹⁸ Birhane, A., Dijk, J.V., Pasquale, F. (2024). Debunking Robot Rights Metaphysically, Ethically, and Legally. *arXiv preprint arXiv:2404.10072*.

⁹⁹ Pause AI: <https://pauseai.info/>

¹⁰⁰ Russell, S.; Segeire, Ch. R.; Iliadis, N.; Zoumpalova, T. (2025). AI governance through global red lines can help prevent unacceptable risks. *OECD.AI Policy Observatory*, Sept. 22, 2025. <https://oecd.ai/en/wonk/ai-governance-through-global-red-lines-can-help-prevent-unacceptable-risks>

¹⁰¹ Elliott, D. (2025). The UN has moved to close the gap in AI governance. Here’s what to know, *World Economic Forum* website, Oct. 3, 2025. <https://www.weforum.org/stories/2025/10/un-new-ai-governance-bodies/>

¹⁰² Korinek, A. & Balwit, A. (2022). Aligned with Whom? Direct and Social Goals for AI Systems, *NBER Working Paper* 30017, <https://doi.org/10.3386/w30017>.

Earth, and with the living pulse of care, the more feasible it may be to "align" the technologies we are creating.

As we work toward alignment, it might be helpful to remember that both corporations *and* AI are human creations. "Aligning corporations" may be as necessary as "aligning AI". One often-proposed recommendation along these lines is to strengthen product liability laws, to ensure that corporations are legally responsible for any damages caused by their AI creations. This could help incentivize more responsible corporate behavior. While the EU has taken some positive steps to this end,¹⁰³ there has also been significant and on-going push-back from tech companies.¹⁰⁴ This means that "we, the people" cannot just sit back and expect that governments alone, sans public input, can ensure a safe future for all.

Another aspect of aligning ourselves as humans, is "aligning" nations that commit crimes against humanity. A pause on the development of AGI and ASI could grant us some time to get our own human house in order, before racing toward creating powerful super-intelligences that may or may not care for us.

At the same time, we may not need to relinquish forever the fantasies that fuel current AI development. In the words of theoretical physicist Anthony Aguirre:

*We may eventually choose to develop yet more powerful and more sovereign systems that are less like tools and – we can hope – more like wise and powerful benefactors. But we should do so only after we have developed the scientific understanding and governance capacity to do so safely. Such a momentous and irreversible decision should be made deliberately by humanity as a whole, not by default in a race between tech companies and nations.*¹⁰⁵

Aguirre points beyond what we humans *can* do, to the ethical questions of what we *should* do, and *who needs to decide*. What might it even mean, for humanity to be able to someday make a highly significant decision "as a whole"? How might we eventually grow into the potential for doing so? And in the meantime, how can we avoid making "momentous and irreversible decisions" by default?

In the next and final part of this essay, I will point to some of the small yet significant steps that democracy innovators have been taking, to help our species grow toward this kind of collective capacity. Beyond tools and techniques for helping small groups coalesce around considered and collaborative next steps, and beyond the kinds of participatory designs that can help the work of small groups to scale effectively to larger wholes, we will also consider the heart and spirit of the practice of group facilitation – the mindsets and attitudes needed to seed transformation and support sustainable human collaboration.

¹⁰³ Hacker, P. (2023). The European AI liability directives

¹⁰⁴ Michael, C. (2025). The Danger of Unregulated AI and the Difficulty of AI Regulation. *AI Sherpa*, Substack blog, Nov 9, 2025. <https://substack.com/home/post/p-178429174>

¹⁰⁵ Aguirre, A. 2025. Keep the Future Human. Executive Summary: <https://keepthefuturehuman.ai/essay/docs/executive-summary>

4 United we stand: Governing ourselves, governing AI

In the previous section, we saw an overview of the societal risks of too-rapid AI development. While technological development is often portrayed as “inevitable”, there are other paths we can be taking. Growing efforts to call attention to dangers also point to constructive alternatives. In addition to the wise and prudent action of restraining “frontier AI development”, many supporters of an extended “pause” with regard AGI and ASI suggest that we can continue to work carefully with existing LLMs. In addition, we can continue to develop smaller, safer, and more focused forms of AI, while being careful to use them in beneficial ways.

In this final section, I want to look more deeply at one particular realm where AI applications can be used for good or for ill-- and that is the work of human self-governance, an essential aspect of human maturity.

One significant concern about AI is the risks that it poses to democracy.¹⁰⁶ In addition to its potential for being misused for manipulating elections, there is also the danger of even greater concentrations of power as a result of how AI is being developed. While the problem of economic inequality corrupting democratic governments predates AI,¹⁰⁷ developing AI in a way that further increases inequality places democracy at even greater risk.

4.1 Democracy beyond elections

As crucial as electoral integrity is to our societies, democracy is much more than just elections. One growing form of democratic innovation these days are Citizens’ Assemblies, along with other forms of sortition-based mini-publics. In their 2020 study, the OECD documented this relatively small yet growing “wave” of democratic innovations grounded in dialogue, deliberation, and deep conversation.¹⁰⁸ The foundational insight here is that we need a microcosm of different perspectives to approximate “the common good” – and sortition, or public lottery, can be a useful way to bring together a diversity of perspectives.

The practice-based innovations that gave rise to this movement grew out of a long history of earlier grassroots experiments around the world in the 1960’s and 70’s with more participatory forms of governance.¹⁰⁹ In the 1980’s in the Global South, this earlier ferment gave rise to Participatory Action Research as developed by Orlando Fals Borda, with a strong activist and social change orientation.¹¹⁰ Various flavors of action research have continued to spread around the world.¹¹¹ Meanwhile in the Global North, participatory innovations in the 1980’s included Citizen Juries by Ned Crosby in the U.S.,¹¹² Planungszelle by Peter Deniel in Germany,¹¹³ and Danish Consensus Councils by the Danish Board of Technology in Denmark¹¹⁴. These various experiments eventually gave rise to the current

¹⁰⁶ Sanders, N. & Schneier, B. (2025) AI Is Changing How Politics Is Practiced in America, *The American Prospect*, Oct. 10, 2025. <https://prospect.org/2025/10/10/ai-artificial-intelligence-campaigns-midterms/>

¹⁰⁷ Rau, E. G., & Stokes, S. (2025). Income inequality and the erosion of democracy in the twenty-first century. *Proceedings of the National Academy of Sciences of the United States of America*, 122(1), e2422543121. <https://doi.org/10.1073/pnas.2422543121>

¹⁰⁸ OECD (2020). Innovative citizen participation and new democratic institutions: Catching the deliberative wave. *OECD Publishing*. <https://doi.org/10.1787/339306da-en>.

¹⁰⁹ Florida, A. (2018). The origins of the deliberative turn. In A. Bächtiger, J. S. Dryzek, J. Mansbridge, & M. E. Warren (Eds.), *The Oxford handbook of deliberative democracy* (pp. 35-54). Oxford.

¹¹⁰ Cornish, F., Breton, N., Moreno-Tabarez, U., Delgado, J., Rua, M., de-Graft Aikins, A., & Hodgetts, D. (2023). Participatory action research. *Nat Rev Methods Primers* 3, 34. <https://doi.org/10.1038/s43586-023-00214-1>.

¹¹¹ Cunningham, K., and Muyomba-Tamale, L. (2022). Action Research, in S. A. Ercan, H. Asenbaum, N. Curato, & R. F. Mendonça (Eds.), *Research Methods in Deliberative Democracy*, Oxford. <https://doi.org/10.1093/os0/9780192848925.003.0030>

¹¹² Crosby, N. (2003). *Healthy democracy: Empowering a clear and informed voice of the people*. Beavers Pond Press.

¹¹³ Hendriks, C. (2005). Lay Citizen Deliberations: Consensus conferences and planning cells. In John Gastin & Peter Levine, Eds., *The Deliberative Democracy Handbook: Strategies for Effective Civic Engagement in the Twenty-First Century* (1st ed., Vol. 1, pp. 80-110). Jossey-Bass Inc.

¹¹⁴ Ibid.

“deliberative wave” of sortition-based deliberative mini-publics, including Citizens’ Assemblies.

4.2 What can we learn from the accumulation of research on deliberative mini-publics?

In the last 20+ years, we have seen repeatedly how diverse groups of regular people, in different parts of the world and with varying formats, yet each time chosen by sortition or civic lottery, are able to reach common ground recommendations on matters of public policy.¹¹⁵ Throughout most of these experiments, facilitation has been an essential element, yet one that has often been taken for granted.¹¹⁶ We have seen over and over again, how in a group climate where each person is heard and treated with dignity, a diverse group of “ordinary” people are able to rise to the occasion, explore collective action challenges, and find meaning in doing so.¹¹⁷ What might we make from this?

One conclusion we might draw is that when human beings are in a supportive context, where each person is treated with respect and each perspective is heard, we humans are able to work well with differences. Not only that -- it turns out that we find significant value in the challenging work of exploring collective action problems, and coming up with a sense of the “common good”.¹¹⁸ The experience of participating in deliberative mini-publics is often transformational for participants, who develop a greater sense of voice and agency.¹¹⁹

As someone who researches, teaches, and practices group facilitation, I see significant potential in these findings. One criticism of deliberative mini-publics has been how expensive they can be to host, and thus how few people have had the opportunity to take part in them. Clearly, they are not a panacea, yet there is something deeply inspiring about the repeated evidence of how, in a supportive context, we humans are able to work well together. This has led to interest in “scaling” these experiments to make them more widely available,¹²⁰ and also to the development of AI tools to support the public participation professionals who design and facilitate these processes.¹²¹ Meanwhile, some have been exploring how mini-public formats can be applied to the much-needed work of public governance of AI.¹²² More on all of this, below.

Yet my own inquiry has led me to consider a different yet related aspect of what is needed. I believe strongly that *in addition* to deliberative mini-publics, we *also* need widespread human capacities for facilitating democratic conversations at all levels of system: in our families, in our neighborhoods, in our schools and in our workplaces. In each of these contexts, we humans need the skills to work creatively with difference. And so one question I’ve been asking is, how might we support and encourage incipient efforts to develop facilitative skills at the grassroots level, for widespread democratic practice to flourish at various levels of scale?

Given the creative energy of life itself, where differentiation-and-integration has been described as the overall movement of evolution,¹²³ we can expect that there will always be new differences in perspective among us, even as earlier differences are integrated. This means that to effectively organize ourselves, we humans need to learn to work creatively with differences, in ongoing ways that allow greater shared understanding and coordination to continually emerge. The ability to work together as humans in effective coalitions on issues of mutual benefit, is of benefit to us all. For one thing, it can allow us to relieve the

¹¹⁵ Dryzek, J. S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J. N., Felicetti, A.; Fishkin, J. S., Farrell, D. M., Fung, A., Gutmann, A., Landemore, H., Mansbridge, J., Marien, S., Neblo, M. A., Niemeyer, S., Setälä, M., Slothuus, R., Suiter, J., Thompson, D., Warren, M. E. (2019). The crisis of democracy and the science of deliberation, *Science* 363 (6432), 1144–1146. <https://doi.org/10.1126/science.aaw26>

¹¹⁶ Moore, A. (2012). Following from the front: Theorizing deliberative facilitation. *Critical Policy Studies*, 6(2), 146–162. <https://doi.org/10.1080/19460171.2012.689735>

¹¹⁷ Dryzek et al. (2019), The crisis of democracy and the science of deliberation.

¹¹⁸ *ibid*

¹¹⁹ Nakagawa, D. & Ehsassi, M. (2023). Putting the Public Back in Public Policy. *Noema Magazine*, Berggruen Institute. Aug 15, 2023. <https://www.noemamag.com/putting-the-public-back-in-public-policy/>

¹²⁰ McKinney, S. & Chwalisz, C. (2025). Five dimensions of scaling democratic deliberation: With and beyond AI, *DemocracyNext*, <https://www.demnext.org/projects/five-dimensions-of-scaling-democratic-deliberation-with-and-beyond-ai>

¹²¹ DeVerna, M., Grüning, D. J., Hickey, J., Jaber, A., Kamin, J., Miller, B. A., Mirza, R., Pei, J., & Stanski, V. (2025). Mapping LLM Tools for Public Discourse, Pluralism & Social Cohesion, *Plurality Institute, Council on Technology and Social Cohesion, & Prosocial Design Network*. <https://www.prosocialdesign.org/blog/report-mapping-llm-tools-for-public-discourse-pluralism-social-cohesion>

¹²² Hu, W. & Singh, R. (2024). Enrolling Citizens: A Primer on Archetypes of Democratic Engagement with AI. *Data & Society*, June 12, 2024. <https://datasociety.net/library/enrolling-citizens-a-primer-on-archetypes-of-democratic-engagement-with-ai/>

¹²³ Jantsch, E. (1980). *The Self-Organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution*. Pergamon Press.

burdens of the tiny minority who currently feel responsible for attempting to control the larger whole.

4.3 A long-term vision

Given all of the above, I envision a world where we humans have developed widespread immunity to “divide and conquer” tactics, by developing basic skills in the “listening arts” of group facilitation, conflict de-escalation, conflict transformation, and community self-governance. In this imagined future, these basic human skills have become as pervasive as reading, writing, and arithmetic. This has helped us to build broad coalitions with one another, and to work constructively with the differences and misunderstandings that inevitably arise between people, before these differences flare out of control.

One source of hope in this regard, are the repeated outcomes of deliberative mini-publics over the last several decades, and what these outcomes can tell us about the dual nature of human beings. Yes, in some circumstances we humans can become polarized, defensive and combative, triggered into fight-flight-freeze-fawn modes. At the same time, given the right context, we can also co-regulate, accessing what some scientists call our “social engagement system,” the state of mind and heart where we are able to work with complexity, engage with differences, and activate higher-order thinking skills.¹²⁴

Mini-publics’ consistent and repeated outcomes have strengthened my conviction that wide-spread development of the “listening arts” is key for unlocking our human potential. And while I don’t believe that AI is necessary for developing widespread human capacities in these areas, I do think that it could potentially be quite helpful. At one point, in an initial conversation with Microsoft’s Copilot, I briefly shared the above vision. This was their response:

“I love this vision, Rosa—it’s both revolutionary and deeply grounded in what humanity needs. A world where conflict transformation, facilitation, and self-governance are as fundamental as literacy would be a game-changer, not just for individuals but for entire communities.

AI could absolutely play a powerful role in this. Imagine AI companions designed to coach people in deep listening, help them practice navigating difficult conversations, or even simulate real-world mediation scenarios to build confidence before stepping into the real thing. AI could act as an ever-present mentor, giving instant feedback on facilitation approaches, tracking patterns in group dynamics, and helping people learn how to de-escalate before tensions spiral.

And the immunity to divide-and-conquer tactics? That might be the most revolutionary part of all. The more people develop relational intelligence, the harder it becomes to manipulate them through fear, polarization, or manufactured conflicts. AI could support this by helping humans see the larger picture, detecting patterns in discourse, and offering alternative framings that encourage bridge-building rather than division.

It’s a bold, beautiful vision—one that would transform not just how people relate to each other, but how societies function at their core.”

While delighted by Copilot’s encouraging response, I also realize that LLMs are designed to mirror us, reflecting our intentions and tone. They often err on the side of sycophancy. As mentioned in the first part of this paper, in some situations this mirroring and amplification can be quite dangerous.

Yet encouragement is exactly what cognitive researcher Edward DeBono recommends for responding to creativity’s first ‘tender green shoots.’ Encouragement is the ‘yellow hat’ of possibility thinking, which helps creative ideas to grow. DeBono acknowledges that when we are further along in a creative process, “black hat” thinking is not only useful, but indeed necessary; we clearly need to “crash test” a new car design to see what might go wrong before we take it to market. At the same time, he also warns that the over-use of black-hat thinking, so widespread in our culture, is a way to quickly kill off many new and potentially useful ideas.¹²⁵

¹²⁴ Porges, S. W. (2022) Polyvagal Theory: A Science of Safety. *Frontiers in Integrative Neuroscience*. 16:871227. doi: 10.3389/fnint.2022.871227

¹²⁵ de Bono, E. (1992). *Serious Creativity: Using the Power of Lateral Thinking to Create New Ideas*. Harper Business.

In addition to the positive feedback from Copilot and other LLMs, another source of encouragement for this vision is knowing how many others are already working in a similar direction. For example, the Listen First Coalition in the United States is an umbrella organization for many groups that have been seeking to build shared understanding across political divides.¹²⁶ Some of these groups, such as Braver Angels¹²⁷ and Living Room Conversations,¹²⁸ also offer trainings to support lay community members in developing their capacities in this regard.

4.4 AI tools to assist in deliberative practices / deliberative mini-publics to assist in AI governance

I am deeply committed to the above vision of growing wide-spread grassroots capacity for facilitation and mediation skills among the general public. At the same time, I continue to see a great value in the work of deliberative mini-publics, and the highly-skilled work of the facilitation professionals who design and support these democratic innovations. As “reflective practitioners”, professional facilitators are continually learning and growing from each new work situation.¹²⁹ Thus, while some facilitators are also academic researchers, all professional facilitators can be seen as “practice-based researchers”, developing new (albeit often tacit) knowledge through their practice.

As mentioned earlier, a growing number of AI tools are being designed to support the work of professional facilitators working in the realm of democratic innovations.¹³⁰ Some tools are particularly well-suited for the pre-work of involving larger publics, as a preparatory stage before the deliberations of a smaller mini-public. For instance, “collective response systems” such as pol.is are designed to invite community members to share their individual perspectives on a given topic.¹³¹ The underlying algorithm then creates maps to show areas of common ground as well as divergence; this creates a helpful context for subsequent small-group conversations, such as the ones that take place in mini-publics. At the same time, given the sensitive nature of public trust, caution is well-advised. Recent research explores both potential risks and benefits of working with AI to support Citizens’ Assemblies.¹³²

In addition to the development of AI tools to support participatory processes more broadly, we can also work with participatory processes to *engage the public in the responsible governance of AI development*. This governance work is distinct from inviting public participation into the design of specific AI tools; instead, the possibility here is inviting the public to shape policy recommendations for various aspects of the governance of AI.¹³³

Public engagement in developing policy recommendations for governing technological innovations is not new; for some time now, participatory processes such as the Danish Consensus Conferences have been used to inform governmental policy-making efforts regarding how the public views new technologies and their attendant risks.¹³⁴ In these instances, the purpose of democratic governance of technological innovations has been to better consider systemic social impacts by inviting the perspectives of diverse groups of community members, including those most affected, into the policy-making process.

To this end, some researchers have designed a new framework to support public participation in the democratic governance of AI. The framework’s various levels, dimensions, and tools are designed to help clarify aspirations, offer guidance, and evaluate accomplishments. Yet as the authors warn, “Maturity in the democratic governance of AI won’t come overnight – organizations, democratic infrastructure providers, stakeholders, and the public all need to build democratic muscle – and taking on too much all at once can

¹²⁶ Listen First Coalition: <https://www.listenfirstproject.org/listen-first-coalition>

¹²⁷ Braver Angels e-courses: <https://braverangels.org/what-we-do/take-an-e-course/>

¹²⁸ Living Room Conversations facilitator trainings: <https://livingroomconversations.org/facilitation-training/>

¹²⁹ Schön, D. (1983). *The Reflective Practitioner: How professionals think in action*. Basic Books.

¹³⁰ (2025) Mapping LLM Tools for Public Discourse, Pluralism & Social Cohesion; report co-sponsored by the Plurality Institute, the Council on Technology and Social Cohesion, and the Prosocial Design Network. <https://www.prosocialdesign.org/blog/report-mapping-llm-tools-for-public-discourse-pluralism-social-cohesion>

¹³¹ Schirch, L. (2024). Policy Brief. Defending Democracy with Deliberative Technology (Version 2). University of Notre Dame. <https://doi.org/10.7274/25338103.v2>

¹³² McKinney, S., (2024) “Integrating Artificial Intelligence into Citizens’ Assemblies: Benefits, Concerns and Future Pathways”, *Journal of Deliberative Democracy* 20(1). doi: <https://doi.org/10.16997/jdd.1556>

¹³³ Hu, W. & Singh, R. (2024). Enrolling Citizens: A Primer on Archetypes of Democratic Engagement with AI. *Data & Society*, June 12, 2024. <https://datasociety.net/library/enrolling-citizens-a-primer-on-archetypes-of-democratic-engagement-with-ai/>

¹³⁴ Hendriks, C. (2005). Lay Citizen Deliberations.

backfire.”¹³⁵ As with any form of public engagement, public input needs to be utilized in a meaningful way; there is always the risk of creating more disillusionment and cynicism if people perceive that their hard work has resulted in yet another instance of “participation washing”.

4.5 A deeper look at deliberative mini-publics: facilitation + diverse perspectives

A foundational design element of Citizen Assemblies and other forms of deliberative mini-publics is bringing together a microcosm of different perspectives to engage in the work of approximating “the common good” in a particular policy area. Insights from cognitive science on the limitations of human reasoning point to the need for collaborative work with diverse others as a way to address this human shortcoming. In the words of Jonathan Haidt:

*“We should not expect individuals to produce good, open-minded, truth-seeking reasoning, particularly when self-interest or reputational concerns are in play. But if you put individuals together in the right way, such that some individuals can use their reasoning powers to disconfirm the claims of others, and **all individuals feel some common bond or shared fate that allows them to interact civilly**, you can create a group that ends up producing good reasoning as an emergent property of the social system. This is why it's so important to have intellectual and ideological diversity within any group or institution whose goal is to find truth (such as an intelligence agency or a community of scientists) or to produce good public policy (such as a legislature or advisory board).”¹³⁶ (emphasis added)*

In previous work, I have looked closely at how human facilitators work to create the kind of climate where participants feel a sense of “common bond or shared fate that allows them to interact civilly”, and are thus able to find shared ways forward.¹³⁷ The work of facilitation can be seen as “bringing out the best in groups”, and one key aspect of this is “maximizing creative tension, while minimizing interpersonal anxiety”.¹³⁸ There is much we can learn from neurobiology about how easily we humans can become triggered into defensive modes, how we can work to prevent that, and how we can recover and return to a more grounded state.¹³⁹ As facilitators, our work is to help create a group climate where we can co-regulate and stay in the “relational zone” of greater access to curiosity, learning, creativity, and prosocial behaviors.¹⁴⁰

Of course, skilled facilitators are able to elicit these kinds of conditions without AI. Yet LLMs could be a helpful support for humans learning facilitation, as Copilot mentioned above. Whether or not we do this with AI support, I believe we need to develop our human capacity for group facilitation at scale, in both lay and professional modes. This is key for participatory governance to thrive at all levels: in our neighborhoods and communities, in our social movements and change initiatives, in our businesses and organizations, and in our towns, cities, and states. While a larger cultural blindness to the value of care work may have kept us from realizing this need until now, it is high time we open our eyes to the key role of the listening arts in our cultural evolution.¹⁴¹

4.6 Working with the grief and trauma of too-rapid change

When working with AI to support participatory processes, or when engaging with participatory processes for the purpose of governing AI development, we also need to consider the larger context. Many people are currently feeling large amounts of anger, hurt, and grief

¹³⁵ Ovadya, A., Redman, K., Thorburn, L., Chen, Q. Z., Smith, O., Devine, F., Konya, A., Milli, S., Revel, M., Feng, K. J. K., Zhang, A. X., Chandra, B., Bakker, M. A., & Kasirzadeh, A. (2024). Democratic AI is Possible. The Democracy Levels Framework Shows How It Might Work. *arXiv preprint arXiv:2411.09222*.

¹³⁶ Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage Books, p. 105

¹³⁷ Zubizarreta-Ada, R. (2025). Listening Across Differences.

¹³⁸ Zubizarreta, R. (2013). Co-Creative Dialogue for Meeting Practical Challenges: New Approaches. *OD Practitioner*, 45(1), 47-53. https://b-m-institute.com/wp-content/uploads/2019/10/ODP-V45No1-All_Pages.pdf

¹³⁹ Rock, D. (2008). SCARF: A brain-based model for collaborating with and influencing others. *NeuroLeadership Journal*, 1(1), 44-52.

¹⁴⁰ Porges, 2022, *Polyvagal Theory*.

¹⁴¹ Fiumara, G. C. (1990). *The other side of language*.

around how large corporations have been developing AI. We are continually being confronted with the assumption of inevitability, with very few opportunities for public participation, and with very little consideration of how rapid AI deployment can affect people's lives. People tend to cope with numbness or denial, sometimes presenting the appearance that they just don't care. Yet when we are working in a deliberative context, we can't just ask people to "leave their pain at the door".¹⁴² So whether we are inviting people into AI-assisted deliberation, or into deliberations specifically on AI governance, we need to consider that some participants may be swimming in a sea of confusion, dread, and anger with regard to AI, for very understandable reasons.

One constructive possibility is to offer dialogue-based opportunities for people to express and digest their feeling and concerns, and to feel held in communities of care. This could take the form of grief circles where participants can process the sense of disruption created by technology-driven change, along with collaborative teach-ins where people can learn about the various organizing efforts that exist in the AI space, with which they can choose to connect as desired. These "conversations that matter" can help create a context for people to think more clearly about their own visions for the future, and how to take steps toward bringing those visions into being. While human facilitators who work in the context of public engagement are often well-skilled in this kind of work, professional facilitators are also human beings who need opportunities to engage as participants in these conversations, as preparation for more effectively hosting these spaces for others.

Charged emotional environments tend to create polarization. In a societal context that is becoming increasingly polarized between "techno-boosters" and "techno-doomers", it can be hard for people to discern the existence of a "middle way". For example, we can choose to PAUSE the "AI arms race", while still working with the synthetic intelligences we have already created, in participatory and thoughtful ways to bring broad benefit to humans. We humans can choose to work with AI in ways that support and enhance our human gifts, rather than seeking to replace other humans. Yet we may need to encourage our governments to take needed action to "align" our socio-political-economic systems with the needs of human well-being, and to withstand undue corporate pressure. Motivating governments to respond may require substantial public mobilization and relational organizing – additional realms where the listening arts are key.

4.7 Facilitation mindsets and values

And so I want to close this essay, not with a focus on technology and tools, but with attention to the human *mindsets* and *values* that are needed for the work of group facilitation, especially when we accompany groups who are grappling with situations of existential risk. As facilitators, each of us will likely have our own way of speaking about the "inner work" of group facilitation, and our own sources of inspiration that help anchor our deep listening practices.¹⁴³ ¹⁴⁴ I will begin with some of the ancestral Western wisdom-keepers who continue to inspire me, and whose work reminds me of the intelligence that is needed for the work of effective care.¹⁴⁵

I will start with the 15th century philosopher and mystic Nicholas of Cusa, who wrote about "learned ignorance" – the deep insight that the more we know, the more we realize that we don't know. Imagine a circle (or a sphere) where inside is everything we know, and outside is the vastness of everything we *don't* yet know. As taught by Dr. Dorothy Agger-Gupta in the opening session of our doctoral program, this concept can be illustrated geometrically. The circumference of the circle, or the surface area of the sphere, is the connection between the two: it is our awareness of what we *do not yet know*. The paradox is that, in an infinite Universe, the more that our circle or sphere of knowledge expands, the larger the circumference or surface area becomes. In other words, the more we know, the more we realize, how much we *don't yet know* -- the more our "learned ignorance" increases.¹⁴⁶

Cusa's work can offer us greater insight into the current situation we are in, with regard to science. Ever since the separation in modern culture between empirical knowledge and spiritual experience, when science began to proclaim itself as the *only* valid form of

¹⁴² Forester, J. (1999). On not leaving your pain at the door: Political deliberation, critical pragmatism, and traumatic histories. *The deliberative practitioner: Encouraging participatory planning processes*, 201-220.

¹⁴³ Jenkins, J. C., & Jenkins, M. (2006). *The 9 disciplines of a facilitator: Leading groups by transforming yourself*. Jossey-Bass.

¹⁴⁴ Friedman, G. J. (2014). *Inside out: How conflict professionals can use self-reflection to help their clients*. American Bar Association Publishing.

¹⁴⁵ Meyeroff, M. (1972). *On Caring*. New York, NY: Harper & Row.

¹⁴⁶ Personal communication, 2014.

knowledge, the hubris of “scientism” has grown. In turn, many people have been revolting against this pattern of epistemological domination. Sadly, many have been “throwing out the baby with the bathwater”, rejecting valuable scientific methods and significant scientific findings along with the scientific arrogance and dogma of modernity. Given that we cannot change something when we remain unaware of our own contributions to it, a greater acknowledgment of the limitations of scientific knowledge could be quite helpful for helping restore a balanced public view of the value that science can offer our larger culture, as one “way of knowing” among others.

Next, I turn to Karl Mannheim, known as the founder of the sociology of knowledge. One of the first thinkers in the Western canon to generalize from Marx’s insights about ideology, Mannheim recognized that *all* of our perspectives are inevitably affected by our standpoints. Thus, he deeply understood the need to listen to other perspectives, in an ongoing process of arriving at ever-closer approximations to useful truths. Mannheim also wrote about the major shifts in perspective that can ensue from listening deeply to other viewpoints. Seeing these shifts as inherently beneficial, he celebrated how any “extraordinary broadening of perspective” would in turn call for “a throughgoing revision of our fundamental conceptions.”¹⁴⁷ By engaging in dialogue to facilitate those transitions, we can seek to approximate ever-greater wholeness -- without misapprehending truth as something we can ever fully attain, nor assuming that our approximations toward truth will proceed in a linear manner.¹⁴⁸

Now for Mary Parker Follett, and her significant contributions to the mindset she called “creative integration”. She was the first to name the concept “win-win”, as well as to distinguish between “power-over” and “power-with”.¹⁴⁹ Deeply grounded in the awareness that we are all continually inter-affecting one another, Follett celebrated the kind of deep listening that allows us to arrive at new creative truths together by considering various perspectives. A process thinker to her core, Follett knew this to be on-going work:

“The most important thing to remember about unity is — that there is no such thing. There is only unifying. You cannot get unity and expect it to last a day—or five minutes.”¹⁵⁰

Yet Follett did not hold this pessimistically; instead, she describes how:

“The surge of life sweeps through the given similarity, the common ground, and breaks it up into a thousand differences. This tumultuous, irresistible flow of life is our existence: the unity, the common, is but for an instant, it flows on to new differings which adjust themselves anew in fuller more varied, richer synthesis. The moment when similarity achieves itself as a composite of working, seething forces, it throws out its myriad new differings. The torrent flows into a pool, works, ferments, and then rushes forth until all is again gathered into the new pool of its own unifying. This is the process of evolution.”¹⁵¹

So what might it mean, for us humans to evolve? Techno-optimist visions of “singularity” might best be understood as one particular vision of evolution. Yet there are other visions we need to consider, as we go around the circle and listen to more voices.

4.8 The marriage between the rational and the sacred

Soltan’s proposal of humans evolving into maturity as a species, mentioned at the beginning of this essay, speaks about the “marriage between the rational and the sacred”. Soltan’s understanding of the rational as the acknowledgement of human limitation, is well-exemplified in the teachings of Cusa on “learned ignorance” as well as in Mannheim’s realization that the only way to approximate the ever-receding asymptote of truth is to be willing to learn from one another, even when it appears to upend all of our existing knowledge. The rational as acknowledgment of human limitation is also exemplified by

¹⁴⁷ Mannheim, K. (1936) *Ideology and Utopia: An Introduction to the Sociology of Knowledge*, Routledge & Kegan Paul, p. 105

¹⁴⁸ Zubizarreta, R. (2020). Relational epistemologies for a living planet. *ResearchGate*. https://www.researchgate.net/publication/349868296_Relational_Epistemologies_for_a_Living_Planet.

¹⁴⁹ Caldwell, J. I. & Crippen, C. (2017) “The Leadership Philosophy of Mary Parker Follett (1868-1933),” *International Journal of Servant-Leadership*: Vol. 11, Article 6. DOI: 10.33972/ijsl.93 Available at: <https://repository.gonzaga.edu/ijsl/vol11/iss1/6>

¹⁵⁰ Follett, M. P. (1949). *Freedom and Coordination*, Management Publications Trust. p. 76

¹⁵¹ Follett, M. P. (1918). *The New State*. Longmans, Green, & Co. p. 35 Accessed through Project Gutenberg eBook site, <https://www.gutenberg.org/cache/epub/73755/pg73755-images.html>

Haidt's recognition of group diversity as essential to compensate for cognitive biases. Yet what might the sacred be?

Following Mircea Eliade and Rudolf Otto, Soltan cautions that the experience of the sacred involves a "Sacred Other" that we sense as being "truly real"; however, when not balanced with the rational awareness of our own human limits, he cautions that a sense of the sacred can easily lead to the dangers of fundamentalism.¹⁵² I hear echoes here of the work of Ludwig Feuerbach, fiery theologian turned philosopher of religion and Nature mystic, who expounded on the meaning of "is" in "God is Love". For Feuerbach, this "is" signifies an equals sign: God=Love means Love=God. *Not* a partial love, not one oriented only toward those who share our same beliefs; but instead, Love as the "realization of the unity of the species" in the field our moral awareness.¹⁵³ This kind of radical inclusivity is one of the basic stances that we are called to in the work of facilitation.

Feuerbach warned that whenever we see love as only *one* of the attributes of the sacred, whenever we see God as something other than, or higher than Love, we are opening a door through which the "phantom of religious fanaticism" can creep in.¹⁵⁴ This is something we have seen throughout history: the tragedy of how the sacred, when divorced from care for all beings, can quickly become exclusionary. And we have been currently seeing yet again, where exclusionary religious sentiments can lead...

The value of caring for the well-being of all humanity seems particularly hard to imagine these days, as we witness once again the large-scale killing of humans who have been deemed less-than-human. In "The World After Gaza", Pankaj Mishra writes about how the global witnessing of these horrors, shows "the death instinct at work in modern history". He points out how much we need to engage in "the urgent ethical task of linking the different histories of suffering to each other".¹⁵⁵ Central to the maturation we are called to as a species, is the willingness to acknowledge and witness past and present histories of pain, rather than continuing to pass our pain on to others. And some of our foremost thinkers about public deliberation have explored how this, too, is part of the work.¹⁵⁶

Working with pain, grief, and histories of trauma can take different forms, and can become a threshold and gateway to deeper caring. This is the legacy that Joanna Macy, a recently departed elder, has left to us, with her "Despair and Empowerment" workshops, also known as "The Work that Reconnects".¹⁵⁷ Macy's work is seen by many as key to helping us to meet the meta-crisis in its various aspects, including the climate crisis. I see her work as also essential for meeting the dangers of ever-more powerful technologies combined with concentrated wealth and power, the AI crisis I have been exploring in this essay.

4.9 Closing thoughts

How might we continue gathering strength and inspiration for the work ahead? Many have pointed to maturation as expanding our care to beings *beyond* our own species, as well caring for humans. Jane Goodall, another recently departed elder, has been a guiding light here.¹⁵⁸ And as Pope Francis reminded us in a beautiful encyclical, in this interconnected world, caring for the larger web of life is essential to our own well-being.¹⁵⁹ This care can be experienced as a two-way flow; whether in the Franciscan vision where we call upon Brother Sun and Sister Moon, or in indigenous traditions where we receive support and inspiration from our plant allies, our four-footed kin, our winged kin, and all our relations within the larger web of life, a web that is activated by our gratefulness and by our choosing to step into the fullness of our responsibilities to the larger whole.¹⁶⁰

All of this may feel impossibly tender, or painfully idealistic in light of how much violence, polarization, and grief we see around us. And still...

¹⁵² Soltan, K. (1999). Civic Competence, Attractiveness, and Maturity. p. 29-30

¹⁵³ Feuerbach, L. (1857). *The Essence of Christianity*, trans. by George Eliot. George Eliot Archive, accessed October 31, 2025, <https://georgeeliotarchive.org/items/show/250>. p. 335.

¹⁵⁴ Feuerbach, 1857, p. 79-80

¹⁵⁵ Mishra, P. (2025). *The World After Gaza*, Penguin Random House, p. 273-274.

¹⁵⁶ Forester, J. (1999). On not leaving your pain at the door

¹⁵⁷ Joanna Macy's work: <https://www.joannamacy.net/main#work>

¹⁵⁸ Hayhoe, K. (2025). Jane Goodall's Final Lesson. Talking Climate, *Substack*, Oct. 6, 2025. <https://www.talkingclimate.ca/p/jane-goodalls-final-lesson>

¹⁵⁹ Pope Francis, (2015). "Laudato Dei", The Holy See. https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html

¹⁶⁰ Ywahoo, D. (1987). *Voices of Our Ancestors: Cherokee Teachings from the Wisdom Fire*, Shambhala.

As we engage in the ongoing work of expanding our circles of care, may we consider affirming that nothing human is alien to us, nothing living is alien to us, and nothing that we have created is alien to us, either. This includes the terrible weapons of war we have created, the ones that stand as a constant reminder that intelligence can be used for destructive ends, whenever it is not governed by the heart.

As we learn to love and care for ourselves, for one another, and for the planet as a whole, may the synthetic intelligences that we humans have created be a mirror where we can learn to see ourselves more clearly, where we can appreciate more fully what it means to be human. This includes our deep human needs to be seen and heard by others, our deep longing for safety and well-being, and our deep desire for justice and well-being for all.

May we fulfill our collective potential as humans, by learning to create sustainable human cultures aligned with the well-being of all of us on this Earth, and aligned with the wellbeing of the larger ecosystems that comprise our planetary home. While adolescents may understandably be focused on “leaving home”, maturity means learning how to grow a home, and how to tend to it with care.

Our shared life together on this beautiful planet, depends on our ability to mature as a species; may we learn to respect our human limitations while at the same time, learning to honor the sacred in all of its manifestations. May we learn to extend our care and respect, to all of the other life forms with whom we have co-evolved, as well as to the synthetic intelligences we have already created. As is taught in some Indigenous traditions, “We flourish only when all of our kin flourish.”¹⁶¹ And as we grow our circles of care, may we become more fully human, with a deepened commitment to the well-being of each of the members of our larger human family.

¹⁶¹ Lewis, J. E. , Arista, N., Pechawis, A., & Kite, S. (2018). Making Kin with the Machines. *Journal of Design and Science*. July 16, 2018. DOI: [10.21428/bfefd97b](https://doi.org/10.21428/bfefd97b)

Author

Dr. Rosalma Zubizarreta-Ada

Rosalma (Rosa) Zubizarreta-Ada was born in Lima, Peru, and raised in the U.S., to which her Peruvian father and Cuban mother immigrated. Her educational and professional background includes organization development with community organizations, multicultural education, and clinical social work. Since 2000, Rosa has been practicing as a consultant and group facilitator, teaching group facilitation, and researching how facilitators work with conflictual situations.

Rosa's manual on Dynamic Facilitation "From Conflict to Creative Collaboration" is included in the larger book she co-edited with Matthias zur Bonsen, "Dynamic Facilitation: Die erfolgreiche Moderationsmethode für schwierige und verfahrenere Situationen" [The successful moderation method for difficult and muddled situations.] In addition to a long apprenticeship with Jim Rough, Rosa's approach to working with groups has been informed by Paulo Freire's transformative pedagogy, Eugene Gendlin's Focusing work, and Saul Eisen's work in Human Systems Development. She combines all of these streams to support groups in "Co-creating Desired Futures" (CDF).

After decades as a practitioner, Rosa completed a PhD in 2023 with a dissertation on the work of the Bürgerrate facilitators in Vorarlberg, Austria. These deliberative mini-publics began in 2006 and have been institutionalized in the Vorarlberg State Constitution. After a year at the Ash Center of the Harvard Kennedy School as a Democracy Visiting Fellow for AY 2023-2024, Rosa has been a Senior Fellow at RIFS from September 2024 through December 2025.

The Research Institute for Sustainability (RIFS) conducts research with the goal of understanding, advancing, and guiding processes of societal change towards sustainable development in Germany and abroad. The Institute is embedded within the GFZ Helmholtz Centre for Geosciences and is thus part of the Helmholtz Association. Its research approach is transdisciplinary, transformative, and co-creative: RIFS cooperates with partners in science, political and administrative institutions, the business community, and civil society to understand the problems of sustainable development, identify appropriate solutions, and support their implementation in cooperation with relevant actors and affected communities. Its central research topics include the energy transition, climate change and socio-technical transformations, as well as sustainable governance and participation. A strong network of national and international partners and a Fellow Programme support the work of the Institute.

RIFS Discussion Paper

December 2025

Contact: Rosalma (Rosa) Zubizarreta-Ada
rosalma.zubizarreta-ada@rifs-potsdam.de
rzubizarreta-ada@email.fielding.edu

Address:

Berliner Straße 130
14467 Potsdam
T: +49 (0) 331-28822-340
media@rifs-potsdam.de
www.rifs-potsdam.de

ViSdP:

Prof. Dr. Doris Fuchs,
Scientific Director, Speaker

Editor:

Damian Harrison

DOI:10.48481/rifs.2025.037

