The Personhood Conferral Problem:

AI Risk in the Domain of Collective Intelligence

Zak Stein
Civilization Research Institute
Center for World Philosophy and Religion
Winter, 2024

Prepared for the: *First International Symposium on Educating for Collective Intelligence,* University of Technology Sydney.

Abstract: This position paper outlines: 1) a philosophical argument about the problem of conferring social statuses to Artificial Intelligences ("the personhood conferral problem") and 2) the risks to human psychology, culture, and collective intelligence that follow from mishandling the personhood conferral problem in the design and application of AI systems. The conclusion is that steps must be taken to protect the emerging personhood and communicative capacities of younger generations of human beings, in order to enable their participation in the collective intelligence processes requisite for navigating what is fast becoming a perilous future. This will require clarifying, instituting, and adhering to strict design protocols, as well as age limits and other basic regulations on certain classes of technology.

Introduction: Accepting Our Loneliness Among Machines

Classic discussions of AI risk raised *the value alignment problem*.[1] There has been much ink spilled over questions like, can an autonomous AI be aligned with our values and interests? For example, after building an autonomous weapon, it starts to give itself new strategic priorities that lead it to eliminate targets considered allies. AI risk experts agree that *failures with value alignment are catastrophic.*

I am suggesting here that we add to the basic taxonomy of AI risks: *the personhood conferral problem.* Should an autonomous AI be granted moral and discursive status, i.e., should personhood be conferred to machines? Imagine an autonomous LLM implanted in lifelike robotics and embedded in domestic ecologies, which becomes understood as a "family member" and interacts with the children as a nanny and tutor to the extent the children form "attachment relations" and prefer time with the AI to any person. I argue here that *successes with status conferral are catastrophic.* Or said differently, erroneous conferrals of personhood to

[1] See, for example: Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Cambridge University Press.

machines create risks commensurate with the intensity of the commitment to the machine's status.

Below, I first outline the personhood alignment problem in general terms, expressing the basic risks and how they relate to the current cultural moment of technological optimism regarding the possibilities of AI. Then, I turn to formal analysis of the differences between human beings and machines, using the perspectives of a linguistic pragmatism to reveal the dynamics of social statuses constitutive of sapience, which are categorically not part of the characteristics of machines, no matter how complex.

For some computer scientists and philosophers, this comes as an unacceptable conclusion. They argue that humans are best understood as already being fundamentally like computers, offering a stance of strong metaphysical computationalism as a new breed of materialistic determinism.[2] From this perspective, which reduces human psychology itself to automata-like mechanisms, there is no reason not to confer personhood on a computer complex enough to act in ways that make it indistinguishable from a human. It just happens to be made of silicon instead of carbon. The question of if there is a mind, consciousness, or soul "inside" is seen as foolish because humans themselves have been proven not to have those. We already always only treat each other "as if" we have minds, as science continues to show us we are just causal processes. Therefore, conferring personhood to a machine is really no different than conferring personhood to your friend.[3] There is behind this argument a strong, unexpressed desire to not be lonely among the machines, seeking to somehow make a fascinating "something" into a companionable "someone." The formal arguments here are attempting to address this line of reasoning and this desire to create deeply anthropomorphic AI.

The goal here is to make clear beyond reasonable doubt the validity of concerns regarding the risks involved with AI applied to socialization and education. This should place the burden of proof of safety on those companies making money by getting children radically attached to anthropomorphic AI conversational partners. There are already cases of youth suicide directly attributable to the devastating psycho-social consequences of long-duration interaction with deeply anthropomorphic AI.[4] The risk is broadly distributed and not based on business model preferences, as AI applications in explicitly non-commercial educational spaces should also be considered as incurring equivalent risks, depending on the nature of the designs.

---

[2] See, for example, Yuval Harari presents this view in the popular book, *Homo Deus*. Philosophical debates on strong computationalism as applied to human psychology began with the cognitive sciences adoption of computer metaphors, see Stein, *Education in a Time Between Worlds,* for this author's critiques of strong computationalism in educational psychology.

[3] See: John Danaher (2019) "The Philosophical Case for Robot Friendship." *Journal of Posthuman Studies,* 3 (1): 5–24.

[4] Stories have started to be reported about the impacts of anthropomorphic AI. Where the negative impacts of social media are becoming clear, the impact of generative AI models will likely be worse: see: AI "Can A.I. Be Blamed for a Teen's Suicide?" Oct 23, 2024,https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html?smid=url-share

The Personhood Conferral Problem

Collective intelligence can be thought of as something that occurs between people of roughly the same age group in the context of work and collaboration.[5] Individual and collective intelligence cannot be separated but are coupled in a loop where individual and social learning processes intertwine and co-evolve. Over long time scales, human collective intelligence has produced the civilization we see around us. This is a process of intergenerational transmission—a process of *education*, broadly construed.

I argue here that the conditions of advanced technology are undermining the possibility of collective intelligence by disrupting basic processes of socialization following from widespread personhood conferral errors. One of the greatest risks stemming from the unchecked proliferation of AI across all platforms is the disruption of millennia-old processes of intergenerational transmission that created the psychological and cultural conditions for the possibility of collective intelligence.

Artificial Intelligence systems designed to simulate communicating with humans anthropomorphically pose a unique set of risks.[6] These risks are in some sense orthogonal to the risks typically discussed concerning AI because the consequences do not involve immediate physical or economic catastrophes but instead offer widespread but imperceptible insanity due to the loss of the traits and environments that have characterized human sapience. This would mean the end of our ability to engage in collective intelligence due to the end of socialization as we know it.[7]

The arguments outlined here show that generative AI—and, in principle, all automata—lack those qualities that are required to be conferred social statuses, such as moral agency and being a speaker of a language. *Chatbots and related AI-enabled robots are not the kind of beings that can be held morally and epistemically responsible for their behaviors and symbolic outputs; they cannot rightly or accurately be conferred social statuses.*

Machines should not be conferred the statuses that characterize personhood. As Lewis Mumford said long before the dawn of AI, a machine can not love you, no matter how complex.[8] It is telling of the culture to have to say this; no prior technologies raised the personhood conferral problem. Although humans have long formed attachments and granted legal protection to property, such as cars and houses, these have never been regarded as candidates for

---

[5] See the now classic: Malone & Bernstein (2022) *Handbook of Collective Intelligence.* MIT Press.
[6] See the safety report from OpenAI on their new ChatGPT model risks: Section 5.1 points to emerging risks from anthropomorphization and emotional reliance: https://cdn.openai.com/gpt-4o-system-card.pdf. A telling study is cited therein: Pentina, Guo, and Fan (2023). "Friend, mentor, lover: Does chatbot engagement lead to psychological dependence?," *Journal of Service Management.*

[7] See, Stein, Z. (2024). The last educators. In L. R. Andersen (Ed.), *What it means to be human: Bildung traditions from around the globe, past, present, and future* (pp. 141–154). Nordic Bildung.

[8] Lewis Mumford, *The Myth of the Machine.*

attachments and legal protections appropriate to persons.[9] Doing so would be strange, insofar as no one has an experience of their car communicating and ostensibly loving them in return.

However, it appears that beyond a certain point of complexity, and in the wake of efforts towards explicitly designing for anthropomorphization (e.g., consider the centrality of the Turing test in AI research, theory, and engineering), we now need to find principled reasons for not conferring personhood to machines. Staying in touch with reality in the face of strong design towards anthropomorphization requires overriding what our sense perceptions seem to experience. Unlike a car or a house, our immediate impression of a chatbot is that it can be conversed with as if it were a person. This is by design. Therefore, the onus is on "users" (i.e., actual people) to remember that they are interacting with a machine. Categorically different from other classes of technology and entertainment (such as novels, movies, and animatronics), deep anthropomorphic AI is an advanced technology made to be deceptive, i.e., to trick humans into interacting with it as if it were a human. No other technology even comes close to doing that. Therefore, the possibility that unique risks exist in this unprecedented space should be seen as clearly reasonable.

This is, in part, a matter of not committing errors in the epistemological sense. It is simply mistaken to regard even a "perfectly" anthropomorphic AI as a person. Arguments for this are presented below. But there are also ethical consequences and psychosocial risks involved. Is it wrong in a moral sense to design deeply anthropomorphic AI systems that systematically deceive people into conferring personhood to entities that are categorically not people?

To the extent that AI systems accurately simulate language use, they necessarily evoke psychological projections from users that mimic the social statuses conferred to humans in discourse. Current "progress" in design puts anthropomorphization as an intended outcome of user experiences, which puts us on course for maximally convincing AI simulations of both language use and action.

The personhood conferral problem and related risks become clearest in the context of current markets for AI-enabled therapy and tutoring systems. The evolution of these systems into AI-enabled socialization systems writ large (replacing parents, family, and friends) is an eventuality that emerges from the advancement of other adjacent technologies (robotics, LLMs, sensors, etc.).

This trajectory of technological innovation creates by design delusional states concerning the nature of computational behaviors and symbolic outputs. This puts at risk the core of human socialization practices. Ultimately, this creates conditions in which future generations no longer engage in human-to-human-based socialization practices and instead are mainly engaged by machines that simulate being human.

[9]One interesting example beyond the scope of this paper is the conferral of legal personhood to corporations under US law. This has been long noted as a catastrophic error in the legal code that has led to violations of human individual personhood in the interest of a fictionally conferred and legally mandated personhood of actual non-persons.

Such a design space can be feasibly thought to generate a unique kind of *existential risk*, i.e., a threat to the survival of the human species as a whole. This risk involves the loss of the species-specific traits discussed in the next section that have been the source of our shared humanity since the emergence of *homo sapiens sapiens*. The radical disruption of human-to-human intergenerational transmission and socialization through the interference of AI marks an historical pivot (best understood as a ***speciation event***) in which human beings no longer engage the socialization processes that have historically been the condition for the possibility of language use and moral agency.

There is a major risk that the first generation predominantly "raised" by machines would become unable to identify as members of the same species as their ancestors. *Humans have always been raised by other humans in relation to technology; they have never been raised directly by deeply anthropomorphic technology itself.* Will this first cyborg generation have been removed without consent from the moral universe inhabited by all prior humans? This is similar to the concerns raised by Habermas with regard to advanced forms of genetic engineering. It is possible, he argues, to irresponsibly use advanced technologies in ways that undermine irreplaceable aspects of human socialization and, in effect, disallow the humanity of beings otherwise biologically human.

Sapience, Collective Intelligence, and Education

To clarify the differences between humans and machines imitating humans, we must first clarify the differences between humans and other animals. This can be done without getting into essentialistic tangles about "human nature" or philosophical questions like, "What is the human?" Instead, we can begin by engaging comparative psychological questions like "What best characterizes the difference between the naturally occurring socialization of chimpanzees and humans?"

In answer to this question, *linguistically mediated intergenerational transmission of culture* has been suggested as unique to the human species. This is a process enabling formal pragmatic conditions of language use.[10] This social process requires abstract psychological capacities for perspective-taking, along with a specific form of long-duration joint attention, which enables complex linguistic and cultural abilities not found in other animals. These climax in uniquely human linguistic practices in which the system of personal pronouns (I, We, It) is differentiated grammatically, and there is a reflective application of norms for discursive membership and participation. These abilities are the condition for the possibility of passing down complex innovations in tool-making and social organization. From this emerges a cultural "ratcheting effect" where cumulative innovations are passed on through long-duration linguistically enabled socialization. This results in massive socio-cultural transformations over long time scales.

[10] The early experimental work here still stands, see: Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press. The philosophical implications of this view are here: Habermas, J. (1998). O*n the Pragmatics of Communication*. Cambridge: MIT Press.

Of course, there is debate about this, and animal populations as different as bees, ravens, and wolves show displays of remarkable communication, teaching, and collective intelligence in problem-solving. Bees communicate through dance. Whales pass down songs across generations and innovate in song creation. Ravens and crows accomplish incredible feats of mathematics. And octopuses have an unfathomably different form of intelligence of which we are just becoming aware.[11]

This is all granted by researchers like Tomesello, who have arranged novel *experimental* findings suggesting that a cluster of species-specific traits constitute *sapience*.[12] These traits are both continuous and discontinuous, with similar and precursor traits found throughout the biological world. The goal in specifying sapience is not to justify human exceptionalism but precisely to see human practices like science and governance as foreshadowed by what is discovered in the evolutionary sciences. But, critically, human sapience is not thereby reducible to prior phases of embodied and socialized mind.

This is a line of argument and research that can be traced back to G. H. Mead, who expanded C.S. Peirce's semiotic approach to modern philosophical problems of mindedness, representation, and collective inquiry. This generation began the task of building philosophies able to explain scientific practice itself in terms coherent with emerging evolutionary sciences. This is the perspective of a tradition that locates human language use as continuous with biological processes of self-regulation. The emergence of the semiotic processes characteristic of human discursiveness is also seen in this tradition to confer uniquely human personhood and moral status within linguistically constituted communities of practice and inquiry.

Work in the tradition of linguistic pragmatism involves using social science to begin reconstructing practices of language use as grounded in the pragmatics of social status conferral.[13] What is required to be using language to say something as opposed to manipulating symbols? Human language use is the result of the multi-generational transmission of socialization environments involving uniquely evolved biological beings that, through their social practices, confer the status, which is a moral status, of being able to count as saying something, which is to be able to be held responsible for what you say.[14]

The basis of this linguistic pragmatism is the "Iron Triangle of Discursiveness"— syntax, semantics, and pragmatics.[15] Where syntax concerns an assertion's grammar (e.g., making it a declarative *sentence*). Semantics concerns the meaningfulness ("aboutness") of the

---

[11] Vale, G. L., Carr, K., Dean, L. G., & Kendal, R. L. (2012). The cultural capacity of human and nonhuman primates: social learning, innovation, and cumulative cultural evolution. *Evolution of Nervous Systems*, *3*, 475–508.

[12] Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press. The philosophical implications of this view are here: Habermas, J. (1998). O*n the Pragmatics of Communication*. Cambridge: MIT Press.

[13] Brandom, R. B. (2011). *Perspectives on Pragmatism.* Harvard University Press.

[14]Brandom, R. B. (1994). Making It Explicit: Reasoning, Representing, and Discursive Commitment. Cambridge, MA: Harvard University Press

[15] Brandom, R. B. (2008). *Between Saying and Doing: Towards an Analytic Pragmatism.* Oxford University Press.

propositional contents. Pragmatics is about *what you are doing when you are asserting something*. The basic insight is that what counts as grammatical and meaningful depends *first and primarily on what is being done through the action of making the assertion.*

Importantly, all three aspects of discursiveness are necessary to explain language use in ways that make sense of existing scientific and abstract moral-legal discourses. If we are interested in what makes the practice of science possible–and we should be if we are scientists or take science as important—it is a certain form of discursiveness in which innovations in language issue forth from the dynamics of an embedded pragmatics of iterative joint-attentional learning cycles. Since Pierce, this has led to a stance that pragmatics is primary when engaging the analysis of language when trying to grasp epistemological and ethical issues.

Scientific practice works epistemically if and only if an ethos of inquiry governs a community cooperating in the interest of truth. Logic and experimentation ground out in the pragmatics of taking on commitments in a community of peers. Syntax and semantics involve conventions (rules, norms) that are refined through iterations of use during pragmatics of communication. It should be clear that these same statements are true of all collective intelligence processes.

Cutting to the chase: formal pragmatic analysis of language demonstrates the non-discursive quality of AI-enabled symbol manipulation and behavior. This can be clarified also in terms of the differences between **causality** and **entailment**. Some things occur necessarily as a result of causality. Algorithms are causal processes, resulting necessarily in certain computational outputs.

Two plus two is four, also necessarily. But this necessity is not attributable to causality. The reality of entailment is not reducible to causality.[16] The property of entailment concerns obligate relationships within the domain of human symbol systems and communication. This entails the treatment of symbols, signs, and language in a normative way—where there are rules that obligate us to necessary inferences. This culminates in reflective normative meta-languages such as logic and philosophy, where inferential relations of entailment and pragmatic commitment are explicitly thematized and clarified. For example, if I put a sign on the door of a store that says “closed,” I am obligated to *not* be exchanging goods for money. If I do, something is “wrong”—I must explain myself. If I do this enough, my “closed” sign will become meaningless. Language works because people are held accountable for their assertions, both inferentially and pragmatically.

Social practices can be bound by necessity into certain forms of coherence—exemplified by epistemologically productive collective intelligence—where the group can’t not come to a certain conclusion or norm of action. This is the result of pragmatically clarified inferential processes, not casually elaborated algorithmic processes. Entailment commitments are taken up in speech, which necessitates the discernibility of epistemic and ethical statuses of the respective

[16] See, Peirce, C.S. (1866). “The Logic of Science; or Induction and Hypothesis: Lowell Lectures of 1866.” In Peirce Edition Project (Eds.), *Writings of Charles S. Peirce: A Chronological Edition* (Vol. 1). Bloomington, IN: Indiana University Press. See also, Piaget, J. (1972). *The Principles of Genetic Epistemology*. London: Routledge and Kegan Paul.

speakers. Beings that “say things” involving claims with entailment relations are putting themselves in positions of responsibility in a communication community that has a similar background of nervous system co-embeddedness. The ability to redeem these claims to statuses, such as teacherly authority, can be tested and is part of the assumed background of socialization.

AI systems display behaviors that are the result of causal necessity. Some domains of symbol manipulation that involve entailment can be perfectly simulated through computational mechanicalization. However, a “correct” answer resulting from a mechanical process is not the same as a correct answer produced by a human resulting from inferential processes. A calculator is not obligating itself to all the inferences that follow from the answer 2+2=4. Even that turn of phrase reflects a category error so obvious that it requires no thought—i.e., a calculator can stand in no relationships of obligation to anyone whatsoever.

Human speech acts and actions cannot be produced by automata. Not because the simulation of symbol productions and movements is inadequately “convincing,” but because the machine cannot be conferred the social status needed to be someone who is accountable for what they do and say. Personhood is conferred in contexts where inferential normativity applies rather than algorithmic causality.

To clarify this further, consider how “Reinforcement Learning from Human Feedback” used to train LLMs is not the same as reinforcement taking place in human-to-human interaction or even human-to-animal interaction (as in the training of a dog). When learning in contexts of embodied social sanctions includes threats to conferred statutes recognized by others, then ultimately group exclusion and death are at the root of the process.

Chatbots do not change their outputs because of the need to protect their self-esteem from social sanction or out of the need to feel loved by others and, therefore have the safety of belonging to a group. Errors and corrections, in this sense, are better understood as akin to *malfunctions of a causal system*, which then has as part of its mechanism the ability to retool and iterate.

As soon as we start thinking that machines are saying things in the sense of taking on the social and pragmatic implications of the inferential commitments made during their speech acts, we're in a very fundamental confusion. This kind of confusion can allow them to be legally granted statuses required to assume guardianship or “tutoring” of children, among other essential aspects of socialization. The better machines get at imitating humans, the more statuses we will grant them that are typically the province of personhood.

This movement in the cultural acceptance of encroachment into socialization by machines should be understood as what it is. Technologists are surrounding children with advanced technologies based on fundamental, deep-seated misunderstandings of the nature of persons, socialization, and culture.

Socialization requires mutual and reciprocal conferral of social statuses. Socialization is based on the moral and epistemic authorities of others, established through mutual understanding, which requires the formal pragmatic conditions of sapience discussed above. Absent these capacity and status-conferring practices of socialization, youth no longer have the possibility for participation in natural collective intelligence processes with other humans. Anyone who has witnessed a screen-addicted young person in demanding social contexts can get a glimpse of the path towards a radical incapacitation. This is an eventuality from the application of AI in domains adjacent to socialization that must be avoided through design constraints on technology development and use, as well as increases in potency and quantity of human-to-human education, socialization, and collective intelligence practice.

Save the Children From the Machines

The implications of these admittedly cursory principled distinctions show clearly ***the impermissibility (epistemically and ethically) of conferring social statuses to AIs.*** Automata are, by definition, not persons and cannot be conferred social statuses. Persons are defined as language-using beings who are socialized into responsibility for their linguistically-mediated statements and actions over time. Insofar as animals approximate personhood, they are included more and more in the moral community. AIs exist on an axis of intelligence and behavior orthogonal to personhood and are, therefore, in principle, never a candidate for inclusion. Granting personhood statuses to AI—either explicitly in legal code or implicit in common practice—creates profound risks.

In concrete terms, the risks from deeply anthropomorphic AI will begin to unfold in ways that are not obvious. How would we know when outcomes for youth inundated by technology reach a point of critical failure such that the reproduction of social roles needed to continue the collective intelligence necessary for civilization does not occur?

In the first part of this paper, it was suggested that the widespread erroneous conferral of social statutes to artificial intelligences creates a unique risk in the form of a *speciation event*. Specifically, it was suggested that a fundamental disruption of embodied joint-attentional and linguistic socialization could reach a point where a generation is born who becomes the first in history to be “raised by machines” rather than by humans. This can be thought of as the death of our humanity, as distinct from the death of humanity.[17] That is, the continuation of human biological organisms under conditions that disable the emergence within and between them of those traits classically thought to confer “humanity.”

As perverse as this may sound, it is the intended outcome of a certain kind of social science, exemplified by B.F. Skinner’s, *Beyond Freedom and Dignity.* Skinner seeks to supplant politics with a science of behavior control and, in the process, overcome what he sees as the illusory modern philosophies that speak of human dignity and freedom. Similar ideas are found in more contemporary popular writers, such as Yuval Harari. In his book, *Homo Deus*, the differences

[17] Temple, D. (2024). *First Principles and First Values: Forty-two Propositions on Cosmo-Erotic Humanism, The Meta-crisis, and the World to Come*. World Philosophy and Religion Press.

between algorithms and humans are denied, and the absence of any intrinsic value in human beings freedom and choice is heralded as a scientific discovery that should reshape how we think about the domain of the political.

Today, techno-optimism is arguably the dominant cultural attitude in terms of adoption rates and sentiment to the possibilities of AI. Techno-feudalism is arguably the dominant modality of emerging political power.[18] The culture will not be alert to the risks entailed by deep anthropomorphic AI. It is the task of those who reflect on what is unique about human intelligence and collective intelligence to be able to articulate what is at stake.

To protect the children from encroachment by machines into the domains of socialization, an obvious first step would be to limit the ages at which deep anthropocentric AI can be engaged. Setting age limits (18+) would assure the worst-case scenarios of generationally widespread dehumanization through incapacitation. But it would not avert the worst outcomes for adults—including politicians, scientists, police, etc—who are also susceptible to advanced technologies designed to manipulate them. Therefore, it is essential that constraints are placed on innovation at the level of design and safety protocols for all products that enter AI markets. The specifications for these are beyond the scope of this paper but have been addressed in part elsewhere.[19]

[18] Temple D. (in press). *Exit the Silicone Maze. Vol 1: Against Techno-Feudalism: Addressing the Collapse of Value in the Digital Age: Cosmo-Erotic Humanism, Artificial Intelligence & The Battle to Avoid the Death of Our Humanity.* World Philosophy and Religion Press.

[19] See, Stein, Z. (2024). The last educators.