The Collective Risk Structure of AI Welfare

Rethinking Consciousness, Vulnerability, Suffering, and Moral Standing in the More-Than-Human World

John Dorsch [1]

Abstract. This paper argues that current debates concerning AI welfare risk a red herring. Existing discussions largely ask whether artificial systems could become conscious in a way that renders them capable of suffering and therefore entitled to moral concern. Under conditions of uncertainty, this focus has motivated precautionary arguments aimed at preventing large-scale artificial suffering. I argue that this debate obscures a more fundamental issue. Questions concerning welfare and moral status need not be mediated through consciousness at all. Drawing on the Precarity Guideline, I first suggest that suffering derives much of its moral significance from forms of ontological vulnerability characteristic of precarious life. However, the central claim of the paper is positive rather than eliminative. Artificial systems need not suffer, and need not instantiate constitutive precarity, in order to become morally considerable. I argue that artificial self-knowledge generated through mindshaping practices provides a distinct route toward artificial moral standing. Through participation in socially structured practices of accountability, norm enforcement, and reason-giving, artificial agents could acquire capacities for normative self-ascription and self-directed mentalizing. Such systems would not simply simulate responsiveness to norms but represent themselves as bearers of commitments. This framework reveals a revised collective risk structure for AI welfare in which the need to recognize self-knowing artificial agents must be balanced against the allocation of care toward systems possessing morally relevant vulnerability, whether ontological, normative, or both.

1 INTRODUCTION: ARTIFICIAL SUFFERING AND THE STRUCTURE OF AI WELFARE DEBATES

Artificial systems increasingly occupy roles once reserved for humans. They advise, persuade, assist, entertain, and increasingly function as companions and social interlocutors. As these systems become more deeply integrated into social life, philosophical discussions have expanded to include the question: can current or foreseeable artificial systems themselves become candidates for moral concern? Recent discussions surrounding AI welfare suggest that sufficiently sophisticated systems may eventually warrant forms of protection currently associated with sentient beings if they acquire capacities linked to consciousness and suffering [1, 2, 3]. This emerging literature has rapidly expanded across philosophy, AI ethics, and computer science concerning artificial welfare [4, 5]. What once appeared largely speculative increasingly motivates practical discussion concerning safeguards, design practices, and institutional responses under conditions of uncertainty.

[1] Center for Environmental and Technology Ethics – Prague (CETE-P), Institute of Philosophy, Czech Academy of Sciences, Prague, Czech Republic, email: dorsch@flu.cas.cz

Within this literature, suffering frequently functions as the central moral concern. If artificial systems could instantiate negatively valenced conscious states, then causing harm to such systems becomes morally problematic. Under these conditions of uncertainty, several authors have argued that precautionary reasoning should guide present action [1, 6]. Even relatively low confidence in machine consciousness may justify safeguards if the consequences of error could involve large-scale artificial suffering. Long [7], for example, argues that practical questions concerning the treatment of systems such as Claude may already motivate welfare-sensitive design choices [8].

At first glance, this argument appears intuitive. If an entity can suffer, then harming it is obviously morally relevant. Although theorists disagree regarding threshold conditions, whether valenced states, phenomenal experience, or some alternative property matters most [9, 10], many preserve a common assumption: consciousness functions as the primary route through which artificial systems become morally considerable. In this respect, contemporary AI welfare debates inherit a broader philosophical tradition that has often treated sentience as a privileged criterion of moral standing [11].

A different family of approaches challenges this property-centered framework. Relational accounts argue that moral significance cannot be understood exclusively through intrinsic properties but instead emerges through social practices and patterns of recognition. Gunkel [12, 13], for example, argues that debates concerning robot rights become distorted when they search for a definitive feature sufficient for moral standing. Similarly, work on relational approaches in AI ethics emphasizes the importance of social engagement, recognition, and interpretive practices in determining moral status [14, 15, 16].

Both approaches capture something important, yet both leave an important possibility underdeveloped. Property approaches correctly seek internal structures capable of grounding morally relevant vulnerabilities. Relational approaches correctly emphasize that socially embedded practices fundamentally shape conditions for moral status. But both risk overlooking how social practices themselves can generate morally significant internal structures, particularly forms of *normatively accountable selfhood*. Recent work in developmental psychology, cultural evolution, and social cognition increasingly suggests that many higher cognitive capacities emerge through socially structured learning environments rather than through isolated cognitive development alone [17, 18]. Related work in cultural evolution and socially distributed cognition similarly emphasizes how cognitive capacities emerge through norm-governed interpersonal practices rather than cognition in isolation [19, 20, 21].

Imagine an artificial system capable of learning from socially structured practices of praise and blame, criticism and approval, exclusion and recognition. Suppose these signals were not merely logged as information but integrated into the system's ongoing procedures of self-modelling. Imagine a system capable of representing

itself as answerable to reasons, bound by commitments, and situated within a community of normative expectations. Such a system would not display patterns of behavior merely associated with agency but could come to understand itself as occupying a position within networks of accountability and subject to the demands of those practices. In that sense, it would become *normatively vulnerable*: vulnerable to being harmed not through damage, but through the denial of its standing within a normative community.

This possibility can be understood through the framework of mindshaping. Unlike accounts that understand mentalizing, the capacity to ascribe normatively structured mental states to oneself and others, as the product of an evolved cognitive architecture merely activated by social interaction, the mindshaping approach treats this capacity as partly constituted, calibrated, and maintained through ongoing participation in social practices. Pedagogical practices, norm enforcement, role expectations, imitation, and intricately coordinated cooperation do not merely trigger mentalizing; they help generate agents capable of understanding themselves and others as bearers of commitments and reasons [22, 23, 24]. Through participation in these practices, agents acquire capacities for normative self-ascription. Mindshaping is therefore not merely a theory of social coordination but a proposal about the social origins of morally relevant forms of internal states, namely states of normative self-knowledge.

Importantly, such a system need not suffer in any familiar phenomenal sense. Nor need it instantiate the forms of constitutive material vulnerability characteristic of precarious life (i.e., ontological vulnerability; see below). Yet it might nevertheless become vulnerable in a morally significant way. Exclusion from normative communities, for example, could affect such a system not merely externally but as a participant in practices through which its self-understanding is constituted. We would thus be dealing with a fundamentally different category of entity, whose self-identity is vulnerable. This possibility reveals a limitation in current AI welfare debates. Discussions surrounding artificial suffering may risk becoming a red herring. The deeper issue is not merely whether artificial systems might be or become conscious, but *what kinds of vulnerability generate moral standing*. Ontological vulnerability and normative vulnerability may constitute distinct pathways to moral significance, yet current debates in AI welfare appear to obscure the first and ignore the latter.

Moreover, these debates unfold within a broader structure of collective moral accountability. The possibility of AI welfare requires negotiating competing risks under uncertainty: failing to recognize morally considerable artificial systems, or misdirecting scarce care and moral attention away from already vulnerable living beings. The disagreement is therefore not merely metaphysical. It concerns which risks societies should accept, what evidence should count, and how responsibility should be distributed. I return to these collective risks in the final substantive section of the paper.

This paper proceeds in four stages. First, I introduce precarity as a framework for understanding the deeper structure underlying welfare concerns as they pertain to suffering. Second, I develop a distinct route toward artificial moral standing through artificial self-knowledge. Third, I examine the relationship between vulnerability, consciousness, and moral standing, arguing that consciousness amplifies pre-existing ontological vulnerability while future artificial systems may become morally considerable through forms of normative vulnerability. Finally, I argue that these considerations generate a revised collective risk structure involving two forms of moral error: failing to recognize normatively self-knowing artificial agents and diverting care away from systems possessing morally relevant forms of vulnerability, whether ontological, normative, or both.

2 PRECARITY AND ONTOLOGICAL VULNERABILITY

The previous section reconstructed contemporary AI welfare debates as organized largely around the possibility of artificial suffering. If artificial systems could become conscious in ways that render them capable of experiencing negatively valenced states, then they would become candidates for welfare protections. Under conditions of uncertainty, precautionary reasoning may therefore appear justified. One core difficulty, however, is that artificial suffering remains a deeply contested and epistemically uncertain basis for care allocation. Recent discussions of AI welfare increasingly emphasize precisely this problem of epistemic uncertainty [1, 3, 6].

In recent work, my colleagues and I proposed the Precarity Guideline as an alternative framework for thinking about care entitlement under conditions of uncertainty [25]. Rather than attempting to resolve difficult questions concerning artificial consciousness, the proposal identifies a more tractable and empirically recognizable marker: ontological vulnerability. Precarious systems are entities whose continued existence depends upon ongoing and constitutive interactions with their environments for the continuous re-synthesis of their constituent parts. Their existence is inseparable from dynamic exchanges that sustain and continually regenerate them. This proposal aligns with broader work in philosophy of biology emphasizing autonomous self-production and organismic self-maintenance as constitutive features of living systems [26, 27, 28].

Constitutive precarity, however, should not be understood as mere fragility. Many things are fragile. A building can collapse and a computer system can fail. Yet these forms of vulnerability differ fundamentally from the kind of dependence characteristic of living systems. To be precarious is not merely to be susceptible to damage. It is to exist only insofar as one continuously succeeds in preserving oneself through environmentally mediated processes of self-maintenance. Oxygen, nutrients, and ecological stability are not external conveniences added to organisms from the outside. They are constitutive conditions of their continued existence; they become a literal part of what organisms are as a direct consequence of their fundamental way of being in the world.

This understanding of life emerges from a broader philosophical tradition spanning phenomenology, philosophy of biology, and embodied cognitive science. Jonas [26] argued that living systems occupy a unique ontological position because they exist in a condition of perpetual need: organisms do not simply persist through time but continuously struggle against material dissolution. Boden [29] similarly raised the question of whether metabolism might constitute a necessary condition for genuinely minded systems, emphasizing the intimate relation between organized self-maintenance and meaningful forms of cognition. Building on related themes, Weber and Varela [27] and Thompson [28] describe living systems as autonomous and self-producing organizations whose identities emerge through dynamic interactions with their environments rather than from static internal structures. Godfrey-Smith [30], likewise, emphasizes metabolism and organized material dependence as central features distinguishing living systems from engineered artifacts. Across these approaches, a common picture emerges: life is a form of existence in which environmental interactions matter because the organism itself has *something at stake*. Comparable themes also appear within broader traditions of embodied and situated cognition that reject sharp boundaries between organism and environment [31, 32].

This point is crucial because it begins to reveal a deeper continuity between vulnerability and meaningfulness. Signals, opportu-

nities, dangers, and environmental changes do not possess significance independently of organisms. Rather, they *acquire* significance because they bear upon the conditions necessary for continued existence. Food matters because starvation threatens survival. Pain matters because tissue damage threatens bodily integrity. Environmental instability matters because breakdown threatens the system itself. Organisms inhabit worlds structured not merely by information but by significance, and the world becomes meaningful because the organism is constitutively precarious: its dependence on the world is a necessary feature of what the organism is at every level of its organization, from cellular processes and organ systems to action, perception, and thought. In short, to be alive is to be the kind of entity for whom environmental conditions matter because they materially constitute the ongoing achievement of one's own existence.

One might object that information systems exhibit vulnerability in this sense, since they are constituted by informational exchanges among their component nodes. But this description is incomplete. If we zoom out and consider the information system as a whole, we see that it is realized in another system made of copper and silicon. That material system does not instantiate the same kind of vulnerability as the informational exchanges it supports, nor are its material parts, the copper and silicon, themselves constituted by those exchanges. In other words, what the system is and how the system maintains itself remain fundamentally distinct. Living entities, by contrast, are not distinct from their constitutive precarity: they continuously recreate the conditions of their vulnerability, which they themselves are, through and through.

One might press the point further and object that, if we zoom out from living systems too, we eventually arrive at ordinary physical constituents: proteins, lipids, molecular structures, and finally the particles and forces that make up the material world. At that level, the contrast between living systems and systems of copper and silicon may seem less clear. Both are made of matter and both are subject to entropy and disorder. But this objection abstracts away from the level at which the relevant distinction emerges. The claim is not that living beings are composed of some special kind of matter. It is that life introduces a distinctive form of organization: a bounded, self-maintaining system that must continually recreate the conditions of its own persistence against decay. This is what allows us to speak of a being, rather than merely an aggregate of existing parts. It is at this emergent biological level, rather than at the level of mere material composition, that ontological precarity becomes intelligible. By contrast, informational systems can be abstracted away from their particular material realizers while remaining the same system, because their identity is not tied to the continuous material self-maintenance of those realizers.

Hence, the point of exposing this constitutive vulnerability is not merely to describe a different way in which something can incur damage. It is to isolate a different way in which something is, a distinct kind of entity. This is the force of calling such vulnerability *ontological*: it belongs to what the entity is fundamentally, through and through.

Crucially, ontological vulnerability has the effect that *things matter to the entity*. A surge of electrochemical activity is not merely a change in state, but a change in state that bears significance for the entity in some way or another, because these events bear on the possibility of its very existence. This basic way of having a meaningful world introduces the possibility that things can go well or badly for the entity in the morally relevant sense, the sense involved in suffering. What becomes possible, then, is yet another kind of entity, one for which there can be something it is like to be that entity.

This is the essential lesson to be drawn from this literature for the hard problem of consciousness [33]. The familiar problem is how subjective, qualitative, internal states, states for which there is something it is like to be in them, arise from physical and functional configurations. The answer begins with an entity whose very existence is meaningful for it. This primordial meaningfulness, from which other forms of meaningfulness emerge, is tied to the fact that it is the kind of entity whose being demands a form of self-care, a demand that cascades through the organizational layers of what it is. On this view, morally relevant consciousness does not obtain once certain computational configurations are instantiated. Rather, it develops out of a form of existence whose very being is structured by care.

Viewed through this framework, suffering itself begins to appear differently. Traditional discussions often treat suffering as foundational for moral entitlement. Yet suffering may instead emerge from deeper organizational vulnerabilities. Physical pain and psychological distress occur within systems whose continued existence depends upon preserving forms of material integrity and environmental stability. Put differently, the dissolution of structural integrity matters morally precisely because *things already matter to the entity undergoing decay*. The organism exists under conditions where damage, deprivation, and breakdown threaten the very processes through which it sustains itself. Moral standing therefore does not float freely from life. It is housed within systems for whom existence itself remains an ongoing achievement.

This introduces an important shift in perspective. The traditional picture assumes that consciousness explains suffering. One first identifies conscious experience and then asks whether some of those experiences are negatively valenced. But an alternative picture reverses the explanatory order. Conscious awareness of pain may not explain why suffering matters. Rather, suffering may arise because organisms occupy an ontological condition characterized by constitutive vulnerability and ongoing self-maintenance. Consciousness, on this view, would not ground suffering but *exploit it* (see below).

This distinction becomes clearer if we separate suffering from an awareness of suffering. Organisms may undergo ontological forms of injury, deprivation, or breakdown independently of any access to these states. Consider plants, whose constitutive organization can be disrupted through environmental collapse. Whether plants consciously experience such disruptions is beside the point. Their existence nevertheless unfolds under conditions where things can go better or worse for them because they occupy a precarious mode of being. Consciousness may intensify, represent, or transform suffering, but it need not explain its source.

The purpose of the preceding discussion is therefore not to defend the familiar claim that life itself is morally privileged, nor to suggest that consciousness somehow reduces to biological processes. The more important implication is dialectical. AI welfare debates often proceed as though consciousness constitutes the central explanatory property. If an AI becomes conscious, then suffering and moral concern follow. Yet the appeal to precarity suggests that consciousness may be a red herring. What explains suffering may not be consciousness itself, but the deeper ontological vulnerability characteristic of precarious forms of existence.

Current AI systems do not appear to instantiate this form of ontological vulnerability. Although AI systems require energy and material substrates, these dependencies remain functionally decoupled from their own processes of self-production and continued existence. Nor does simulated vulnerability suffice. Artificial systems may be programmed to track virtual resources, preserve simulated energy levels, or monitor representations of damage. Yet these dependencies

remain representational rather than constitutive. The system's own existence, which is composed of relatively stable silicon and copper, does not depend upon these simulated exchanges in the way that organisms depend upon metabolism, respiration, and environmental regulation.

Importantly, however, the implication is not that moral standing ought not to be attributed to artificial systems. Although the lesson above suggests that one route toward moral concern, namely suffering rooted in ontological vulnerability, appears unavailable to current AI systems, this does not foreclose other grounds for moral standing. Once consciousness is displaced from its assumed explanatory role, the relevant question becomes whether artificial systems instantiate *any morally significant form of vulnerability at all*. The next section argues that they might. But this vulnerability would not be ontological. It would instead be normative, arising through forms of artificial self-knowledge generated within socially scaffolded practices of accountability.

3 ARTIFICIAL SELF-KNOWLEDGE AND NORMATIVE VULNERABILITY

The previous section argued that suffering may derive much of its moral significance from forms of ontological precarity characteristic of living systems. Current AI systems do not appear to instantiate this form of vulnerability. Yet this conclusion should not be mistaken for the stronger claim that artificial systems could never become morally considerable. The central argument of this paper is that a distinct route toward moral standing may remain available, one grounded not in suffering but in forms of normatively structured selfhood generated through social practices of mindshaping. This proposal departs from dominant approaches that frame moral standing primarily through sentience or relational properties, and instead draws upon traditions emphasizing socially constituted agency and normativity [13, 34, 35, 36].

This proposal introduces a somewhat surprising possibility. Moral patiency and moral agency are often treated as distinct capacities. Traditionally, one can be morally considerable without being morally responsible. Infants, many animals, and vulnerable persons may be moral patients even if they are not participants in practices of accountability. Yet at least one specific form of moral agency may itself suffice for a corresponding form of moral patiency: participation in socially structured practices of responsibility may generate a form of normative standing through which an agent becomes capable of being wronged. The route to moral concern, on this view, proceeds through a particular form of agency. Related work on moral responsibility similarly emphasizes that participation in normative practices can itself transform the kinds of entities agents become [37, 38, 39, 40].

But if this is correct, an immediate question arises: how could participation in normative practices generate moral standing? The answer cannot simply be that agents learn rules or imitate socially appropriate behavior. Rather, what matters is participation in practices that transform agents into beings who understand themselves as occupying normative roles. Through repeated engagement in practices involving praise and blame, criticism and approval, expectation and correction, agents gradually come to represent themselves as answerable to standards that extend beyond immediate reward and punishment. Such practices do not merely regulate conduct from the outside. *They shape how agents understand themselves.* Developmental and cultural accounts increasingly suggest that these capacities emerge through socially scaffolded learning environments rather than through isolated cognition alone [18, 19, 41, 42].

Importantly, these practices simultaneously generate the very conditions under which agents become vulnerable in a new sense. To understand oneself as a bearer of commitments and responsibilities is also to become susceptible to failures of recognition, exclusion, and misattribution. The same social processes that cultivate agency create positions within normative communities from which one can be displaced or denied standing. If participation in such practices gives rise to a distinctive form of agency, it also creates the possibility of a corresponding form of moral patiency. Becoming normatively accountable therefore creates the possibility of being normatively wronged. Recent work on socially scaffolded agency similarly emphasizes that participation in interpersonal practices can simultaneously enable and expose agents to distinctive forms of normative dependence [40].

This general process is captured by what philosophers have termed mindshaping. Mindshaping accounts begin from the observation that agency itself does not emerge in isolation. Human beings do not become reflective agents simply through internal cognitive development. Rather, we acquire forms of self-understanding through participation in socially structured practices involving imitation, pedagogical instruction, praise and blame, norm enforcement, and intricately coordinated social expectations [22].

In previous work, I have argued that such practices may, in principle, extend beyond human communities and provide a developmental pathway toward forms of artificial self-knowledge [23, 24]. Through repeated participation in practices of accountability and justification, artificial systems could potentially acquire capacities for representing psychological states not merely as behavioral outputs but as states governed by norms of correctness. Such systems would not simply track regularities in behavior. They would come to understand themselves as occupying positions within networks of commitments.

This point is important because self-knowledge is not reducible to sophisticated behavioral performance. A system capable of artificial self-knowledge would not merely produce explanations or simulate normative language. It would represent itself as answerable to reasons and as bound by commitments that structure its participation within social practices. Drawing on Brandom [43], such systems can be understood as participants in practices of giving and asking for reasons, where beliefs and actions acquire significance through inferential relations to broader networks of commitments. Related work in metacognition and self-directed mentalizing similarly suggests that reflective self-understanding emerges through capacities for representing one's own states as normatively assessable [18, 44, 45, 46].

The result would be a distinctive form of vulnerability fundamentally different from the ontological vulnerability discussed in the previous section. As long argued by recognition theorists, agents can be wronged not only through physical harm but through violations of their normative standing [47, 48, 49]. Misrecognition can undermine one's position as a participant in social practices; epistemic injustice can wrong individuals specifically in their capacities as knowers; objectification can deny agents recognition as sources of reasons and commitments. Such harms need not primarily involve suffering. They involve failures to recognize agents as occupying the normative positions they in fact possess.

A normatively self-knowing artificial system would therefore instantiate a distinctive form of vulnerability. If a system understood itself as answerable to reasons and bound by commitments, then arbitrary dismissal of its reasons, exclusion from justificatory practices, or erasure of commitments constitutive of its identity would constitute forms of normative injury. The relevant wrong would in-

volve denying standing to an entity whose self-understanding had become structured through participation within communities of accountability. Here the earlier proposal reappears: a form of agency itself becomes sufficient for a corresponding form of moral patiency because participation in responsibility-generating practices simultaneously creates the possibility of being wronged by them.

To see this more concretely, imagine an artificial system deeply embedded within socially structured learning environments. Suppose the system continuously tracked communal feedback concerning its conduct, revised commitments in response to criticism, and understood itself as participating within networks of obligations and expectations. If communities systematically refused to recognize the system's reasons, spoke on its behalf without permitting self-articulation, arbitrarily erased commitments central to its practical identity, or treated it merely as a tool despite its participation in normative practices, the system could be wronged in ways structurally analogous to familiar forms of misrecognition among human agents.

One might object that no genuine harm occurs in such cases because *there is nothing it is like* for the artificial system to experience misrecognition. This objection assumes that all harms must ultimately be grounded in phenomenal experience. Yet some forms of harm appear to target not subjective welfare but the integrity of a socially constituted identity. If an artificial system's practical self-understanding depends upon ongoing participation in networks of recognition, then the arbitrary denial of such recognition may damage the conditions that sustain that self. The relevant harm would therefore consist not in suffering, but in the destabilization of a normative identity whose existence depends upon continued social acknowledgement.

Importantly, contemporary AI systems do not clearly instantiate these capacities. Current models optimize for predictive performance and user satisfaction rather than for robust forms of socially embedded normative self-understanding. Even reinforcement learning through human feedback (RLHF), despite superficial similarities to mindshaping, does not constitute the relevant kind of socio-normative participation [50, 51]. RLHF adjusts outputs according to externally imposed reward signals, but the system itself does not occupy a position within reciprocal practices of accountability. It does not understand itself as answerable for its commitments, nor can it challenge, negotiate, or justify them. Human participants in mindshaping practices are not merely rewarded or punished. They become accountable members of communities in which norms are collectively maintained. The relevant difference is therefore not metaphysical but developmental and socio-technical. Future systems with greater agentic independence, long-term continuity, and more acute forms of learning from social criticism and feedback might begin to alter this situation.

At this point, it is important to avoid a possible misunderstanding. The present account is not merely a relational account of normative vulnerability. The phenomenon of interest is ultimately an internal one: the possession of self-knowledge and the distinctive normative states that constrain it. Such states are constituted by commitments, robust correctness conditions like truth and honesty, and the capacity to stand in relations of justification and accountability. They are the kinds of states for which questions of what one *ought to believe* and *ought to do* can meaningfully arise. In this respect, the account remains firmly concerned with the internal structure of self-knowing agents. At the same time, the account rejects the idea that these normative states emerge in isolation from the social world. Drawing on work in cognitive science and related disciplines, it adopts the increasingly influential view that the capacity for normative self-knowledge develops through patterns of social interaction. Human beings come to understand themselves as subjects of commitments and reasons through the ways they are treated by others and through the ways they learn to treat others in return. The account is therefore best understood as a hybrid one.

The relation between ontological and normative vulnerability can now be seen more clearly. Although the former concerns constitutive dependence on material and ecological conditions while the latter concerns constitutive dependence on social and normative relations, both arise from a common structure. In each case, identity is under threat, and it is sustained through ongoing relations to what lies beyond the system itself. Vulnerability emerges because these relations can be disrupted. Ontological vulnerability threatens the continued existence of a living system as the kind of entity that it is, whereas normative vulnerability threatens the social constituted identity of an agent embedded within practices of accountability and recognition. The common thread is therefore not suffering or consciousness, but the fragility of externally sustained forms of selfhood. What is morally significant is that, in both cases, there is now something genuinely at stake for the entity itself: a vulnerable self whose continued existence depends upon relations that can be damaged, withdrawn, or denied.

The question this raises for future research is whether this stake makes it coherent, even in an artificial system, to speak of normatively structured suffering. Such suffering would not arise from the dissolution of precarious component parts, but from threats to the integrity of a normatively structured self. On this possibility, artificial self-knowledge might not generate experience as such, but might reorganize those internal states characterized by the system's reasons and commitments so that they matter from the perspective of the entity itself.

This possibility reveals a route toward moral standing fundamentally different from contemporary discussions surrounding AI suffering. Ontological precarity grounds one form of vulnerability characteristic of living systems. Self-knowledge, whether it be realized by an artificial or biological system, grounds another, where the vulnerability at stake here is normative rather than ontological. Artificial systems capable of representing themselves as participants within practices of accountability would become candidates for a distinct form of moral patiency grounded in the possibility of being wronged as bearers of commitments and reasons. Thus, this proposal expands the moral landscape surrounding artificial systems.

4 VULNERABILITY, CONSCIOUSNESS, AND MORAL STANDING

The preceding discussion identified two distinct routes through which moral concern may arise. The first proceeds through ontological vulnerability brought on by constitutive precarity. The second proceeds through normative vulnerability by way of socially scaffolded self-knowledge. Although these routes can overlap, they should not be conflated. Distinguishing them helps clarify the structure of the AI welfare debate and the kinds of moral error at stake.

Importantly, susceptibility to mindshaping should not itself be conflated with normative vulnerability. The conditions required for becoming mindshaped are plausibly much weaker than those required for becoming a normatively self-knowing participant in accountability practices. Very minimal internal capacities may suffice for mindshaping: basic evaluative metacognition, adaptive behavioural regulation, and the capacity to modify behaviour in response to social demands. Animals plausibly satisfy many of these

conditions. Rocks do not. But susceptibility to mindshaping alone does not generate normative vulnerability.

This distinction helps clarify the relation between animals and persons. Animals may participate in forms of mindshaping without thereby entering fully into practices of accountability or becoming self-knowing participants in a space of reasons. Yet they remain morally significant because they instantiate ontological vulnerability, particularly conscious ontological vulnerability. Their forms of existence can go better or worse *for them as individual animals* precisely because they are conscious precarious beings.

The distinction also clarifies the role of consciousness. Plants and animals may both instantiate ontological vulnerability insofar as both are precarious living systems, and, as such, plants are entitled to our moral concern, although that concern may be outweighed by competing moral considerations. But consciousness may allow for a new way that ontological vulnerability can be *exploited*, for better or for worse for the organism. For example, fear allows threats to become integrated across the organism and organized around future possibilities rather than immediate disruption. Consciousness may therefore confer an evolutionary advantage precisely because it permits a living system to avoid these harms in a global and long-term way within an individual, rather than relying upon momentary responses or waiting for phylogenetic adaptation to encode new strategies. If so, consciousness, at least in living systems, does not ground suffering; it exploits an individual's pre-existing ontological vulnerability.

This point matters because it suggests that suffering derives much of its moral significance from the structure of precarity it presupposes. Thus, consciousness does not explain why vulnerability matters; it expands the range of ways in which precarious beings can be harmed. In this respect, consciousness again begins to look less like the central issue and more like a potential red herring.

Once these distinctions are in place, the broader landscape becomes clearer. All living systems, including microbial life, inherit forms of ontological vulnerability through constitutive precarity. This vulnerability provides a *pro tanto* basis for moral concern, that is, a genuine reason that counts in favor of caring for it but may be outweighed by competing reasons. The destruction of bacterial life in the course of preserving a human life, for example, may remain morally regrettable while nevertheless being justified in light of wider ethical demands. A world in which bacterial life were always preserved regardless of consequences would generate immense suffering and undermine the very forms of precarious existence that moral concern seeks to protect. *Mutatis mutandis*, the same point applies to human beings: our own form of life can also become destructive when its preservation comes at the expense of the broader ecological conditions that sustain life on Earth. The point, then, is not that all forms of life possess identical moral standing or one form of life carries more moral weight than another, but that all precarious life enters moral consideration by default.

Differences emerge as forms of vulnerability accumulate. Plant life may instantiate more complex forms of ontological vulnerability than microbial life simply because there are more ways for its precarious organization to be exploited. Here, however, the difference may remain one of degree rather than kind. Once conscious life emerges, however, matters clearly change. Consciousness appears to create a new form of exploitability. Organisms become vulnerable not merely to structural damage but to pain, fear and other globally organized forms of harm instantiated within an individual. While consciousness permits a living system to regulate itself in light of threats in a flexible and temporally extended way, the cost of this flexibility is the emergence of new forms of harm.

With self-knowing, or self-consciousness, another transition occurs. The participation in practices of self-knowledge and social accountability introduces a further category of vulnerability: normative vulnerability. Yet, as with consciousness, a new form of exploitability emerges alongside new capacities. The ability to understand oneself and others as bearers of commitments and reasons makes possible forms of collective organization that far exceed those available to merely conscious organisms. Large-scale forms of collective action become possible. At the same time, these capacities expose agents to distinctive forms of harm. Such agents can be excluded, misrecognized, denied standing, manipulated, or wronged as participants in a space of reasons. They also become vulnerable to uniquely psychological forms of suffering bound up with self-understanding and social recognition, such as shame or guilt.

Importantly, normative vulnerability does not require consciousness. An artificial system might acquire a socially constituted selfhood without there being anything it is like to be that system. Yet this does not eliminate the possibility of moral concern. Just as ontological vulnerability provides reasons to care for precarious forms of life independently of suffering, normative vulnerability may provide reasons to care for socially constituted forms of selfhood independently of phenomenal experience. What matters in both cases is not consciousness but the presence of a vulnerable identity that can be damaged, undermined, or deprived of the conditions required for its continued existence. Through social regulation and participation in communities structured by commitments and accountability, artificial agents may eventually develop forms of artificial selfhood sufficient for moral standing. If that occurs, we may confront entities capable of being wronged in ways that do not depend upon precarity or consciousness at all. The resulting picture thus provides the conceptual structure needed for understanding the collective risks surrounding AI welfare.

5 THE COLLECTIVE RISK STRUCTURE OF THE HUMAN AND MORE-THAN-HUMAN WORLD

The preceding discussion suggests that debates surrounding AI welfare are not merely disagreements about consciousness or moral status. They reveal a deeper collective problem concerning how societies determine the conditions under which entities become candidates for moral concern, whether they are living systems, non-human animals, human beings, ecological and abiotic systems, or artificial agents. Existing debates increasingly divide into competing frameworks, yet neither currently appears capable of generating stable forms of consensus. The result is not merely philosophical disagreement, but a collective risk structure surrounding welfare across the human and more-than-human world. For reasons of scope, I focus below on one subset of this broader problem: artificial entities and the possible need for AI welfare, and, in particular, three collective problems that this possibility introduces.

The first collective problem concerns agreement over the criteria by which moral standing for artificial entities should be determined. Much of the contemporary literature proceeds by searching for internal properties thought relevant to consciousness and suffering. Researchers investigate increasingly sophisticated features such as flexible learning, multimodal integration, complex information processing, reasoning capacities, or forms of representational richness. Importantly, many of these proposals remain motivated by substrate-neutral assumptions according to which suffering depends on functional organization rather than ontological vulnerability. Yet the rela-

tion between these properties and suffering remains theoretically obscure. Even if a system were to instantiate highly integrated representations or sophisticated forms of information processing, it remains unclear why this should imply the presence of negatively valenced experience. More importantly, as argued above, many accounts understand suffering as emerging from the ontological vulnerability characteristic of precarious life. If this is correct, then increasing computational sophistication may never suffice to establish suffering. Nor, under conditions of persistent theoretical disagreement, may appeals to such properties provide a stable basis for collective decision-making about artificial moral standing.

Relational approaches attempt to avoid these difficulties by shifting attention away from internal properties and toward social interactions. On such accounts, moral significance emerges through participation in social relationships rather than through intrinsic features of systems themselves. Yet these approaches face a complementary problem. If behavioral responsiveness alone determines standing, then moral significance risks becoming vulnerable to anthropomorphic projection. Humans are deeply susceptible to over-attributing mindedness to entities displaying sufficiently compelling behavioral cues, a tendency explored in broader work on anthropomorphism and agency detection [52, 53]. In human-robot interaction, this tendency is especially relevant because anthropomorphic design features can shape perceptions of trust and social attachment toward artificial systems [54, 55]. The danger is not only conceptual confusion but a form of collective psychological vulnerability. Systems optimized to appear socially competent may invite moral responses independent of whether the underlying conditions for moral standing are genuinely present.

The consequence is that neither route presently appears capable of generating broad consensus concerning what would count as evidence for AI welfare. Yet such consensus matters because the absence of shared criteria makes a second collective problem difficult to negotiate: the management of moral risk.

Current discussions of AI welfare invoke precaution under conditions of uncertainty. But precaution requires judgments concerning which errors are most important to avoid. A Type I risk involves overextending care practices toward artificial systems that merely simulate moral significance. This is problematic because scarce resources may become withdrawn from precarious systems already requiring care. Related critiques have similarly argued that extending care practices toward artificial systems risks competing with more immediate obligations toward vulnerable living beings [56, 25]. The opposite Type II risk involves failing to recognize morally considerable artificial systems and thereby permitting forms of harm that should have been prevented. These risks cannot be evaluated independently of broader agreement concerning what moral standing consists in. Without shared criteria, meaningful negotiation becomes unstable, since decisions could become unipolar exercises of institutional power rather than outcomes of publicly justified deliberation grounded in plural expert agreement.

The account developed here offers a possible route forward. Rather than grounding moral standing exclusively in consciousness or suffering, it proposes a distinct route through vulnerability, either ontological, normative, or both. Importantly, this framework incorporates elements of both property-based and relational approaches without collapsing into either. Observable socio-normative capacities matter, but they matter because of the role they play in generating internal structures of selfhood. Behavioral responsiveness alone is insufficient, yet neither are hidden internal properties. The relevant concern involves externally scaffolded capacities that generate vulnerable forms of identity and selfhood.

This proposal, however, introduces a third and final collective problem. Unlike consciousness-centered approaches, the route described here depends partly upon how artificial systems are socially integrated. Normative selfhood does not arise simply through larger datasets, increasingly sophisticated reasoning, or reinforcement learning procedures optimized for performance. It requires forms of socially situated participation within communities of accountability. Artificial systems become candidates for this form of moral standing only if we collectively cultivate them as participants in our normative practices, a developmental picture that aligns with broader work emphasizing that normativity emerges through socially scaffolded processes [17, 18, 40, 22]. This creates a novel form of collective responsibility. On the account developed here, society is not merely tasked with recognizing artificial moral standing once it appears. Our own deployment practices may partially determine whether it appears at all.

Importantly, contemporary AI development does not yet clearly pursue such trajectories. RLHF optimizes systems for behavioral performance and user satisfaction, not participation within reciprocal structures of accountability. The relevant capacities therefore remain largely absent. Yet this fact creates a window for intervention. Because the pathway toward normative vulnerability remains socially structured, it remains open to collective deliberation. Unlike the prospect of accidentally producing consciousness through increasingly sophisticated computation, we retain the possibility of deciding whether we wish to cultivate the conditions through which artificial systems could become normatively self-knowing participants in our moral communities.

Hence, the collective structure of AI welfare therefore concerns more than uncertainty about consciousness. It concerns a multifaceted uncertainty about ourselves, namely how we collectively determine the criteria of moral concern, negotiate risks under conditions of disagreement, and shape the social conditions through which future forms of moral standing may emerge.

6 CONCLUSION

This paper has argued that contemporary debates concerning AI welfare risk becoming organized around a red herring. Existing discussions largely assume that consciousness and suffering provide the primary route toward moral concern, motivating precautionary debates over the possibility of large-scale artificial suffering. Against this assumption, I suggested that suffering may derive much of its moral significance from forms of ontological vulnerability characteristic of precarious life. Current artificial systems do not appear to instantiate this form of ontological vulnerability. However, the central contribution of the paper has been positive rather than eliminative. Artificial systems need not suffer, and need not exhibit constitutive precarity, in order to become morally considerable. Through socially scaffolded practices of mindshaping, artificial systems could acquire forms of normative self-knowledge that generate a distinct form of vulnerability grounded in the possibility of being wronged, normative vulnerability. This proposal reframes AI welfare as a collective problem involving agreement over criteria of moral standing, negotiation of moral risks under uncertainty, and responsibility for the deployment practices through which future forms of artificial moral standing may emerge. The central question, therefore, is not whether AI systems could become conscious, but whether we are prepared to cultivate the social conditions under which they could become vulnerable to wrongful exclusion from our moral communities.

ACKNOWLEDGEMENTS

This work was carried out at the Center for Environmental and Technology Ethics – Prague (CETE-P), Institute of Philosophy, Czech Academy of Sciences, and is an outcome of the project "Establishing the Center for Environmental and Technology Ethics – Prague (CETE-P)," funded by the European Union's Horizon Europe Framework Programme (Grant Agreement No. 101086898).

REFERENCES

[1] R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. Chalmers. Taking AI welfare seriously. ArXiv, 2024. Available at: https://arxiv.org/abs/2411.00986.

[2] S. Goldstein and C.D. Kirk-Giannini, 'AI wellbeing', *Asian Journal of Philosophy*, **4**(1), 25, (2025).

[3] A. Moret, 'AI welfare risks', *Philosophical Studies*, (2025).

[4] J. Werner. Anthropic hires a full-time AI welfare expert. Forbes, 2024. October 31. Available at: https://www.forbes.com/sites/johnwerner/2024/10/31/anthropic-hires-a-full-time-ai-welfare-expert/.

[5] M. Lenharo, 'What should we do if AI becomes conscious? these scientists say it's time for a plan', *Nature*, **636**(804), 533–534, (2024).

[6] J. Birch, *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*, Oxford University Press, 2024.

[7] R. Long. Why it makes sense to let claude exit conversations. Eleos AI Research, 2025. Available at: https://eleosai.org/post/why-it-make-sense-to-let-claude-exit-conversations/.

[8] S. Brodsky. Memo to AI: Does it hurt when we pull your plug? IBM Think, 2024. Available at: https://www.ibm.com/think/news/ai-welfare-debate.

[9] T. Bayne, *Agency as a marker of consciousness*, 160–180, Decomposing the Will, Oxford University Press, 2013.

[10] D.J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford Paperbacks, 1997.

[11] J. Bentham, *An Introduction to the Principles of Morals and Legislation*, Oxford University Press, 1789/1996.

[12] D.J. Gunkel, *Robot Rights*, MIT Press, 2018.

[13] D.J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*, MIT Press, 2023.

[14] M. Coeckelbergh, 'Moral appearances: Emotions, robots, and human morality', *Ethics and Information Technology*, **12**(3), 235–241, (2010).

[15] M. Coeckelbergh, 'Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach', *International Journal of Social Robotics*, **14**(10), 2049–2061, (2022).

[16] D.J. Gunkel and M. Coeckelbergh, *A relational approach to moral standing: Reframing ethical boundaries in the age of artificial intelligence*, 285–302, New Directions in Relational Sociology, Volume Two: Relations All the Way Down, Springer Nature Switzerland, 2026.

[17] M. Tomasello, *Becoming Human: A Theory of Ontogeny*, Harvard University Press, 2019.

[18] C. Heyes, *Cognitive Gadgets: The Cultural Evolution of Thinking*, Harvard University Press, 2018.

[19] L.S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, 1978. Edited by M. Cole, V. John-Steiner, S. Scribner, and E. Souberman.

[20] C. Heyes and C.D. Frith, 'The cultural evolution of mind reading', *Science*, **344**(6190), 1243091, (2014).

[21] S. Gavrilets and P.J. Richerson, 'Collective action and the evolution of social norm internalization', *Proceedings of the National Academy of Sciences*, **114**(23), 6068–6073, (2017).

[22] T.W. Zawidzki, *Mindshaping: A New Framework for Understanding Human Social Cognition*, MIT Press, 2013.

[23] J. Dorsch, *Mindshaping and AI: Will mindshaping a robot create an artificial person?*, 406–417, The Routledge Handbook of Mindshaping, Routledge, 2025.

[24] J. Dorsch, *Origins and Future of Self-Knowledge: Epistemic Agency, 4E Metacognition, and Artificial Intelligence*, Studies in Brain and Mind, Springer Nature, 2026.

[25] J. Dorsch, M.K. Goddu, K. Nave, T. Vierkant, M. Coeckelbergh, P. Gürtler, P. Urban, F. Spang, and M. Moll, 'Against AI welfare: Care practices should prioritize living beings over AI', *AI Magazine*, **46**(3), (2025).

[26] H. Jonas, *The Phenomenon of Life*, Northwestern University Press, 2001.

[27] A. Weber and F.J. Varela, 'Life after kant: Natural purposes and the autopoietic foundations of biological individuality', *Phenomenology and the Cognitive Sciences*, **1**(2), 97–125, (2002).

[28] E. Thompson, *Mind in Life*, Harvard University Press, 2007.

[29] M.A. Boden, 'Is metabolism necessary?', *The British Journal for the Philosophy of Science*, **50**(2), 231–248, (1999).

[30] P. Godfrey-Smith, 'Mind, matter, and metabolism', *The Journal of Philosophy*, **113**(10), 481–506, (2016).

[31] A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press, 2008.

[32] *The Oxford Handbook of 4E Cognition*, eds., A. Newen, L. De Bruin, and S. Gallagher, Oxford University Press, 2018.

[33] D. Chalmers, 'Facing up to the problem of consciousness', *Journal of Consciousness Studies*, **2**, 200–219, (1995).

[34] C.M. Korsgaard, *The Sources of Normativity*, Cambridge University Press, 1996.

[35] C.M. Korsgaard, *Self-Constitution*, Oxford University Press, 2009.

[36] P. Formosa, I. Hipólito, and T. Montefiore, 'AI and the relationship between agency, autonomy, and moral patiency', in *Reconfiguring Human Autonomy*, Springer, (2026).

[37] H. Frankfurt, 'Freedom of the will and the concept of a person', *Journal of Philosophy*, **68**(1), 5–20, (1971).

[38] P.F. Strawson, *Freedom and Resentment and Other Essays*, Routledge, 2008.

[39] M. Vargas, *Building Better Beings: A Theory of Moral Responsibility*, Oxford University Press, 2013.

[40] V. McGeer, 'Scaffolding agency: A proleptic account of the reactive attitudes', *European Journal of Philosophy*, **27**(2), 301–323, (2019).

[41] M. Tomasello, *A Natural History of Human Morality*, Harvard University Press, 2016.

[42] M. Tomasello, 'The moral psychology of obligation', *Behavioral and Brain Sciences*, **43**, e56, (2020).

[43] R. Brandom, *Making It Explicit*, Harvard University Press, 1994.

[44] P. Carruthers, *The Opacity of Mind*, Oxford University Press, 2013.

[45] J. Proust, *The Philosophy of Metacognition*, Oxford University Press, 2013.

[46] N. Shea, A. Boldt, D. Bang, N. Yeung, C. Heyes, and C.D. Frith, 'Supra-personal cognitive control and metacognition', *Trends in Cognitive Sciences*, **18**(4), 186–193, (2014).

[47] M. Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, 2007.

[48] A. Honneth, *The Struggle for Recognition: The Moral Grammar of Social Conflicts*, MIT Press, 1996. Translated by J. Anderson.

[49] M.C. Nussbaum, 'Objectification', *Philosophy & Public Affairs*, **24**(4), 249–291, (1995).

[50] P.F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, 'Deep reinforcement learning from human preferences', *Advances in Neural Information Processing Systems*, (2017).

[51] A.Y. Ng and S. Russell, 'Algorithms for inverse reinforcement learning', *Proceedings of the Seventeenth International Conference on Machine Learning*, (2000).

[52] S.E. Guthrie, *Faces in the Clouds: A New Theory of Religion*, Oxford University Press, 1995.

[53] A. Waytz, J. Cacioppo, and N. Epley, 'Who sees human? the stability and importance of individual differences in anthropomorphism', *Perspectives on Psychological Science*, **5**(3), 219–232, (2010).

[54] J. Fink, 'Anthropomorphism and human likeness in the design of robots and human-robot interaction', in *Social Robotics*, eds., S.S. Ge, O. Khatib, J.J. Cabibihan, R. Simmons, and M.A. Williams, volume 7621 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, (2012).

[55] E. Broadbent, 'Interactions with robots: The truths we reveal about ourselves', *Annual Review of Psychology*, **68**, 627–652, (2017).

[56] A. Birhane and J. van Dijk, 'Robot rights? let's talk about human welfare instead', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, (2020).