PHILADELPHIA JOURNAL OF ARTIFICIAL AND DIGITAL MINDS

A curated review in the Frontiers of Intelligence initiative • Inaugural Issue • The Three Axes • Movement IV of V

WHY THE SAME DISENTANGLING YIELDS OPPOSITE COUNSELS FOR MIND AND AGENCY

Precaution and Non-Deferral

Elan Moritz
Eagles Perch Press • Philadelphia, PA

July 17, 2026

Keywords: precautionary principle; moral uncertainty; sentience; AI governance; moral status; decision theory; agency; mind

About this issue

This is the fourth movement of the inaugural issue of the *Philadelphia Journal of Artificial and Digital Minds*, a five-part sequence building one analytical framework for thinking about artificial minds. The first three movements separated *competence*, *agency*, and *mind* into three independent axes and began to put the separation to work. This movement draws the framework's normative lessons together, and finds them to be, at first sight, opposite: the mind axis counsels *precaution*, the agency axis *non-deferral*. It argues that the opposition is only apparent— the two are one rational principle reading two mirror-image epistemic situations—and completes the framework proper. The final movement then turns the whole apparatus outward, onto a live public debate about the dangers of artificial intelligence. Read in order, the five are one argument.

Abstract

The preceding movements applied a single method—separating a property the vocabulary of "artificial intelligence" fuses into one—to different properties. They isolated *mind* (experience and intentionality) from competence, and *agency* (autonomous action) from both. Each yielded a practical counsel, and the counsels are opposite. For mind the counsel is *precaution*: because a subject might be present and undetected, act as though it may be, and extend protective consideration under uncertainty. For agency the counsel is *non-deferral*: because power is present and undeniable, govern it now, and do not wait on any metaphysical question to resolve. This movement asks why one disentangling produces two opposite disciplines, and answers that the divergence is not ethical but epistemic. Mind and agency differ not in how much they matter but in what we *know* about their presence: the morally relevant property on the mind axis is of *uncertain* presence, while the morally relevant property on the agency axis is of *certain* presence. Precaution is the rational response to a possible-but-undetected morally relevant good; non-deferral is the rational response to a certain-but-unascribed morally relevant power. The two counsels are therefore not in tension; they are the two halves of a single decision-theoretic structure, distinguished by the epistemic status of what the disentangling revealed. The movement formalizes this structure, states the conditions under which each counsel applies, and draws the consequence for how a mature treatment of artificial minds should be organized: not around a single posture toward "AI," but around a property-by-property assignment of postures keyed to what is known about each.

1 One Method, Two Counsels

The same act of disentangling that separated mind from competence, and agency from both, issued in two opposite recommendations—wait-and-protect for mind, act-and-govern for agency. That the identical method should counsel opposite disciplines is the puzzle this movement resolves, and the resolution is that the two properties are in different epistemic situations, not different ethical ones.

The movements of this issue share a single move. The census that anchors the journal holds that the concept of mind must be split into what mind *is* and what it is *made of* (Moritz 2026a). The opening movement extended the split, separating *intelligence*— competence, the ability to achieve goals across environments (Legg and Hutter 2007)—from *mind*, the felt character of experience and the intentional directedness of thought (Moritz 2026b). The second and third movements added a third term, *agency*, the capacity for autonomous action, and showed it independent of the other two: a system can act without being clever and act without being a subject (Moritz 2026e,f).

Each disentangling carried a practical payload, and here is the puzzle. From the separation of mind came a counsel of PRECAUTION: since a competent system might harbor experience we cannot detect, and since wrongly denying moral status to a being that has it is a grave error, one should extend protective consideration wherever the probability of sentience is non-negligible, *even under continuing uncertainty* (Moritz 2026d). From the separation of agency came a counsel of NON-DEFERRAL: since an autonomous system already wields real causal power whether or not anything is home, one should govern that power now, on the causal facts alone, and *not wait* on the metaphysics of machine mind to settle (Moritz 2026e). Wait, in the one case; do not wait, in the other. The same method, the same underlying separation of properties—and opposite marching orders.

A careless reader might take this for inconsistency. It is the reverse: it is the sign that the method is working, because the method separates properties, and different properties can be in different situations that rationally demand different responses. The task of this movement is to make that structure explicit—to show that precaution and non-deferral are not rival philosophies of how to treat artificial systems but complementary responses to two distinct epistemic predicaments, and to state the rule that assigns each response to its case.

2 The Two Postures, Precisely Stated

Both counsels are responses to *uncertainty*, but they respond to uncertainty about different things, and the difference is the whole matter.

PRECAUTION is a response to uncertainty about the *presence of a morally relevant property*. Its canonical form, drawn from environmental and now animal ethics, shifts the burden of proof: where there is a realistic possibility of a serious harm, lack of full certainty shall not be a reason to postpone protective measures (Birch 2024; Sandin 1999). Applied to mind, the uncertain property is sentience: we do not know whether a given system feels, the question may be undecidable by any test now available, and the posture says that this very undecidability, combined with the gravity of the possible error, is a reason to act protectively rather than to withhold action until certainty arrives. Precaution is a discipline for acting well when you *do not know whether the morally relevant thing is there*.

NON-DEFERRAL is a response to uncertainty of a different kind—not about whether a morally relevant property is present, but about how to *ascribe and govern* a property whose presence is not in doubt. On the agency axis the causal power is certain: the system acts, the effects are real, the funds move and the messages send. What is uncertain is elsewhere—in whether

there is a genuine agent to bear responsibility (there is not), in how to route the answerability that the machine cannot hold, in how to keep human control across the joint that autonomy strains (Matthias 2004; Santoni de Sio and Hoven 2018). The posture says: do not let *that* uncertainty—the unresolved metaphysics of machine agency and mind—delay the governance of the power, because the power is present and harmful-capable now, and waiting is itself a choice with consequences. Non-deferral is a discipline for governing well when you *know the morally relevant thing is there* and are tempted to wait on a side-question before acting.

The two uncertainties

Precaution addresses *uncertainty of presence*: is the morally relevant property (a subject, capable of being harmed) there at all? The property might be present and undetected. Error of omission—failing to protect a subject that exists—is the feared mistake. Counsel: act protectively despite not knowing. **Non-deferral** addresses *uncertainty of ascription* around a property whose presence is *certain* (causal power, capable of doing harm): the power is unquestionably there; what is unsettled is who answers for it and how it is bounded. Error of commission—deploying ungoverned power while waiting on a side-question—is the feared mistake. Counsel: govern now, do not wait. The first waits-and-protects because it might be too late to recognize a subject; the second does-not-wait because the power is already loose.

3 Why the Divergence Is Epistemic, Not Ethical

It would be easy to misdiagnose the divergence as a difference in values—as though mind called for a caring, protective ethic and agency for a controlling, managerial one. That reading is wrong, and seeing why is the crux.

The two counsels share an identical ethical core: minimize the expected moral cost of error under uncertainty. Precaution minimizes it by weighting the grave cost of failing to protect a subject that turns out to be real. Non-deferral minimizes it by weighting the grave cost of leaving real power ungoverned while one waits on an irrelevant question. Both are expected-cost-minimizing responses to uncertainty; both are, at bottom, the same rational principle (Steele 2006; Tarsney et al. 2024). What differs is not the principle but the *input*: the epistemic status of the morally relevant property.

The asymmetry that generates everything

On the mind axis, the morally relevant property is the *presence of a subject*, and its presence is *uncertain*—possibly there, undetectable, its existence the open question. On the agency axis, the morally relevant property is the *presence of causal power*, and its presence is *certain*—indisputably there, measurable, its existence not in question at all. Feed "uncertain presence of a grave good" into expected-cost minimization and you get precaution: protect it in case it is real. Feed "certain presence of a grave power" into the same minimization and you get non-deferral: govern it, because it is real and waiting costs. Identical ethics, opposite outputs, because the epistemic inputs are mirror images. The subject might be there and we cannot tell; the power is there and we can.

This is why the divergence is a feature and not a bug. A single posture toward "artificial intelligence"—always cautious, or always permissive—would be a category error of exactly the kind the whole framework was built to expose, because "artificial intelligence" is not one property but a bundle, and the members of the bundle are in different epistemic conditions. The correct

posture is not global but property-indexed: precaution toward the members whose presence is uncertain and whose mistaken denial is grave; non-deferral toward the members whose presence is certain and whose ungoverned exercise is grave. To demand a single attitude toward ”AI” is to insist on treating the bundle as the atom—which is the founding mistake the disentangling corrects.

4 The Third Property, and the Reason It Draws Neither Counsel

The comparison so far has two terms, but the framework has three, and the third—competence, intelligence proper—is instructive precisely because it draws *neither* counsel. Seeing why completes the structure.

Intelligence, uncoupled from agency and mind, is *inert*. A pure oracle's competence, no matter how vast, changes nothing in the world and is no one; it neither acts nor experiences. Its presence may be certain, but the property is not, in itself, morally relevant—competence harms no one and is harmed by no one until it is joined to agency (which lets it act) or to mind (which lets it be wronged). So the intelligence axis calls for neither precaution nor non-deferral, because neither error—failing to protect a subject, leaving power ungoverned—can arise from competence alone. This is not a gap in the account; it is a confirmation of it. The counsels track *moral relevance*, and competence acquires moral relevance only through the other two axes. Intelligence is the axis on which the question ”what posture?” correctly returns ”none, until coupled.”

Research Question

This yields a clean decision rule over the three axes, and stating it exposes the further questions. For any system, and any axis, the posture is fixed by two facts: whether the axis's property is morally relevant (can its presence generate an error of omission or commission?) and, if so, whether its presence is certain or uncertain. Competence: not independently morally relevant; posture, none. Mind: morally relevant (a subject can be wronged), presence uncertain; posture, precaution. Agency: morally relevant (power can do harm), presence certain; posture, non-deferral. The open questions live at the couplings. What posture governs a system high in agency *and* suspected of mind—does the certain power's non-deferral or the uncertain subject's precaution take priority, and do they even conflict, or do they compound into a demand both to govern the power and to protect the possible subject? What happens as an axis changes epistemic status—if a test for machine sentience matured, mind would migrate from uncertain to certain presence, and its posture would shift from precaution to something nearer non-deferral, the way the presence of pain in a familiar animal grounds not precaution but straightforward obligation? The framework predicts that postures are not fixed to axes but to the *epistemic status* of axes, and that improving our detection instruments does not merely refine our knowledge but *changes our duties* by moving a property from the uncertain column to the certain one. Mapping that migration is the research program this comparison opens—and one the later work of this review takes up directly.

5 Two Objections to the Symmetry

"The postures do conflict in the systems that matter most." The hardest case is the system that is both highly agentic and a live candidate for mind—exactly the frontier system the whole framework concerns. Here, one might argue, precaution and non-deferral pull apart: non-deferral says govern and constrain the power firmly, while precaution about a possible subject might counsel restraint in how we constrain it, lest we wrong a being that can be wronged. This is real, but it is not a conflict between the two counsels; it is the ordinary situation of a single entity that falls under two independent duties at once, as a human worker is both a locus of power to be governed and a subject to be respected. That we owe an entity both governance-of-its-power and consideration-of-its-possible-experience is not a contradiction but a conjunction, and the disentangling is what lets us see that two duties are in play rather than collapsing them into one muddled attitude. The counsels compound; they do not cancel.

"Certainty of power is overstated; we are often unsure what a system will do." It is true that we are frequently uncertain about the *specific* actions an autonomous system will take—indeed that unpredictability is the engine of the responsibility gap. But this concedes nothing, because the non-deferral counsel rests on the certainty that the system has *power of some consequential magnitude*, not on the predictability of its particular acts. The uncertainty about which action, precisely, only strengthens the case for governing the power in general, since it is exactly the unpredictable-but-consequential actor that most needs bounding. Certainty that there is grave power to govern is compatible with, and even amplified by, uncertainty about the power's specific exercise. The mind case is the true mirror: there we are uncertain whether the morally relevant property exists *at all*, not merely how it will manifest.

6 Conclusion

Precaution and non-deferral are not two philosophies of AI but two outputs of one rational principle—minimize expected moral error under uncertainty—applied to two properties in mirror-image epistemic conditions. The subject might be present and cannot be detected; the power is present and cannot be denied. Wait-and-protect for the first; act-and-govern for the second; and never mistake the bundle for the atom by demanding a single posture toward "artificial intelligence" as such.

The movements of this issue set out to separate what the phrase "artificial intelligence" fuses: competence, mind, and agency. This one shows that the separation was not merely analytic housekeeping but the precondition of coherent practical guidance, because the three properties call for three different postures, and the differences among the postures are principled. Competence, morally inert until coupled, calls for none. Mind, morally grave but of uncertain presence, calls for precaution—the discipline of protecting what might be there. Agency, morally grave and of certain presence, calls for non-deferral—the discipline of governing what indisputably is there. The counsels look opposite and are in fact complementary: the same expected-cost minimization, reading two mirror-image epistemic situations.

The lesson generalizes past these three properties. Whenever we confront a novel kind of system, the useful questions are not "is it intelligent?" or "should we be worried?" but a disciplined pair asked property by property: *is this property morally relevant, and is its presence certain or uncertain?* The first question routes to whether any posture is owed; the second routes between precaution and non-deferral. This is the mature shape of the inquiry the census began—not a verdict on the minds we are building, and not a single stance toward them, but a method for assigning to each of their separable properties the posture that the epistemic facts about that property rationally require (Moritz 2026a,c). The minds may or may not be there. The powers already are. And competence, the thing we actually know how to build, is by itself neither a

subject to protect nor a power to govern—until we join it to one of the others, at which point the questions this movement has ordered are the questions we will need to have already learned to ask.

With the framework's three axes now separated and their postures assigned, the conceptual work of this issue is complete. What remains is to see the instrument used in the open, on a dispute already underway. The final movement turns from construction to application, taking up a live and heated public argument about the dangers of artificial intelligence and showing how the disentangling of the axes dissolves a confusion at its heart.

Acknowledgement. The author makes extensive use of frontier large language models, as well as local large language models, in the course of this work.

Conflict of interest. The author declares no conflict of interest.

Funding. This work received no external funding.

References

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

Legg, S. and M. Hutter (2007). "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds and Machines* 17.4, pp. 391–444. DOI: 10.1007/s11023-007-9079-x.

Matthias, A. (2004). "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata". In: *Ethics and Information Technology* 6.3, pp. 175–183. DOI: 10.1007/s10676-004-3422-1.

Moritz, E. (2026a). *Conceptions of Mind: A Historical Census*. Frontiers of Intelligence. Philadelphia: Eagles Perch Press.

– (2026b). "The Competence and the Cogito: Why a Mind Need Not Be Intelligent, and an Intelligence Need Not Be a Mind". In: *Philadelphia Journal of Artificial and Digital Minds* 1, pp. 1–8.

– (2026c). "The Dual-Horizon Framework: Toward a Secular Operating System for Civilizational Action". In: *SSRN*. Available at SSRN: https://ssrn.com/abstract=6503099. DOI: 10.2139/ssrn.6503099.

– (2026d). *The Minimal Viable Sentience Problem: Can Frontier LLMs Feel and Suffer?* Philadelphia: Eagles Perch Press.

– (2026e). "The Power Without the Person: Why Delegated Causal Agency Is the Conflation That Cannot Wait". In: *Philadelphia Journal of Artificial and Digital Minds* 1, pp. 19–28.

– (2026f). "The Third Axis: Agency, Competence, and Mind as Three Independent Things". In: *Philadelphia Journal of Artificial and Digital Minds* 1, pp. 9–18.

Sandin, P. (1999). "Dimensions of the Precautionary Principle". In: *Human and Ecological Risk Assessment* 5.5, pp. 889–907. DOI: 10.1080/10807039991289185.

Santoni de Sio, F. and J. van den Hoven (2018). "Meaningful Human Control over Autonomous Systems: A Philosophical Account". In: *Frontiers in Robotics and AI* 5, p. 15. DOI: 10.3389/frobt.2018.00015.

Steele, K. (2006). "The Precautionary Principle: A New Approach to Public Decision-Making?" In: *Law, Probability and Risk* 5.1, pp. 19–31. DOI: 10.1093/lpr/mgl010.

Tarsney, C., T. Thomas, and W. MacAskill (2024). "Moral Decision-Making Under Uncertainty". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Spring 2024. Metaphysics Research Lab, Stanford University.