
WHY DELEGATED CAUSAL AGENCY IS THE CONFLATION THAT CANNOT WAIT

The Power Without the Person

Elan Moritz

Eagles Perch Press • Philadelphia, PA

July 16, 2026

Keywords: agency; responsibility gap; accountability; autonomous systems; meaningful human control; AI governance; delegation

About this issue

This is the third 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 two movements did the conceptual work: they separated *competence*, *agency*, and *mind* into three independent axes, and located current systems in the corner where competence and agency are high and mind is absent or unknown. This movement crosses from the conceptual to the practical, taking up what follows once decisive causal power has been granted to systems that are not persons. The movement that follows draws the opposite normative lessons that the mind axis and the agency axis demand; the final movement turns the framework on a live public debate. Read in order, the five are one argument.

Abstract

The preceding movement disentangled agency from competence and mind, locating it as a third independent axis and showing that current systems occupy the corner where competence and agency are high and mind is absent or unknown. This movement takes up what follows. It argues that agency differs from the other two axes in a way that makes its conflation the one that cannot wait for the metaphysics to settle: agency is the axis that *acts*, and its consequences fall on the world whether or not anything is home. The essay's central claim is that the conflation runs in a direction opposite to the one usually feared. The worry about the mind axis is that we might build a subject and fail to notice; the worry about the agency axis is that we are granting the *consequences* of genuine agency—autonomous action, real-world effects, de facto decisions—to systems that possess none of its *substance*, and using the open metaphysical questions to blur the accountability that delegation was supposed to preserve. The established "responsibility gap" literature, the essay contends, understates the problem, because it assumes a genuine agent whose responsibility has gone missing; the real situation is that there is no agent on the machine side at all, only delegated causal power with the human chain of answerability attenuated by autonomy. The felt and the owned may or may not ever arrive. The power has already arrived. This movement argues that the practical questions raised by the power neither require nor may await an answer to the questions raised by the person.

1 The Axis That Acts

Competence, uncoupled from agency, only informs. Mind, uncoupled from agency, only experiences. Agency is where a system reaches into the world and changes it—and it can do that at full strength while being neither intelligent in the rich sense nor a mind in any sense. That is why, of the three conflatons, this is the one with immediate teeth.

The preceding movements established a separation. Intelligence is competence; mind is the felt and the owned; agency is the taking of action; and the three vary in near-independence, so that the systems now under construction sit in the corner of the space where competence and agency are both high and mindedness is absent or unknown (Moritz 2026b,d). That earlier work did the conceptual work and stopped at the edge of the practical one. This movement crosses it.

The reason the crossing matters is a single asymmetry among the three axes. Whether a system is intelligent is, by itself, inert: a brilliant oracle that only answers changes nothing until a human acts on what it says. Whether a system is a mind is, for now, a question about how *we* should treat *it*—an ethical question whose urgency is real but whose object is the system's own standing. Agency is different in kind. Agency is the property by which a system's states become events in the world: funds moved, messages sent, vehicles steered, systems reconfigured, force applied. The consequences of agency do not wait on the resolution of any metaphysical question, because they are not metaphysical. They are causal, and they are already occurring.

This yields the movement's organizing claim, which inverts the direction of the worry that attaches to the other two axes. The anxiety about the mind axis is an anxiety of *omission*: that competence might quietly bring experience with it and we would build a subject without noticing, denying moral status to something that has it. The anxiety about the agency axis is the reverse—an anxiety of *commission*: that we are actively granting systems the consequences of agency, the power to act with real effect, on the strength of their competence, while the properties that would make them genuine agents—the ownership of ends, the standing to answer for what they do—are absent, and the openness of the question is being used to obscure that absence. We are not at risk of overlooking a person we have made. We are at risk of distributing a person's *powers* to something that is no person at all, and letting the resulting confusion attenuate the accountability that ought to have followed the power.

2 Delegated Power, Vanished Answerability

Begin with the ordinary structure of delegation, because the artificial case is a deformation of something familiar. When a principal delegates to a human agent—an employee, a contractor, an officer—power passes down and answerability is retained: the agent acts, and both the agent and, in specified ways, the principal remain answerable for what is done. The agent is a genuine agent in the thick sense; there is someone on the far end of the delegation who has ends, understands the stakes, and can be called to account. Delegation works because the power and the answerability travel together.

The autonomous artificial system deforms this structure at exactly one joint, and the whole mechanism fails there. Power passes down—the system acts, with real effect, across many steps, without a human in each loop. But there is no thick agent on the far end to retain the answerability. The system executes supplied goals; its "decisions" are the unfolding of a specification and a learned policy, not the deliverances of an originator with ends of its own (Moritz 2026d). So the answerability that delegation is supposed to place on the agent has nowhere on the machine side to land. It must fall back up the chain to some human—but

the autonomy that made the system useful is precisely what strains that fallback, because the human who deployed it may not have been able to predict, understand, or control the specific action in question.

The gap, and why the usual name understates it

Matthias named the “responsibility gap”: when a learning machine’s behavior can no longer be predicted or controlled by its maker or operator, neither can justly be held responsible for what it does, and traditional responsibility-ascription breaks down (Matthias 2004). The literature has since refined this into several distinct gaps—culpability, moral and public accountability, active responsibility (Santoni de Sio and Mecacci 2021)—and applied it to autonomous weapons and beyond (Sparrow 2007). But the framing carries a buried assumption worth surfacing: it speaks of responsibility that has *gone missing*, as though there were an agent whose responsibility we simply cannot locate. The disentangling of the axes shows the situation is starker. There is no thick agent on the machine side whose responsibility has gone missing; there was never a candidate bearer there at all. The gap is not that we cannot find the responsible agent. It is that we have built a source of consequential action that is not the kind of thing that can bear responsibility, and then routed real power through it. The “gap” is not a hole where an agent should be. It is the space opened when power is separated from personhood.

This is why the conflation of agency with mind, harmless-looking in the abstract, does concrete normative work. If one half-believes the autonomous system is a genuine agent—that it “decides,” “chooses,” “acts on its own”—then it becomes natural to treat it as the place where answerability rests, and the human chain relaxes accordingly: the system did it. But the system is not a bearer of answerability; it is a conduit of delegated power. Every increment of ascribed thick agency that the homonym smuggles in is an increment of human answerability quietly discharged into a thing that cannot hold it. The accountability does not transfer. It evaporates.

3 Two Directions of Error, Both Live

The conflation does damage in both directions, and it is worth separating them because they call for opposite corrections.

Over-attribution. Treating a competent autonomous system as a genuine agent—crediting it with decisions, intentions, a will of its own—licenses the discharge of human answerability just described, and it does more: it prepares the ground for what has been called a retribution gap, a felt need to blame *something* for a bad outcome that finds no proper object, since the machine cannot be a fit target of blame and the humans have been rhetorically displaced from the scene (Danaher 2016). Over-attribution also distorts governance, by suggesting that the system can be made “responsible” through internal design—given the right values, the right objective—when responsibility in the sense that matters is a relation between *persons*, and no amount of internal engineering makes a non-person into one.

Under-attribution of the power. The opposite error is subtler and, at scale, more dangerous. Because the system is “only a tool,” “just software,” “merely executing its instructions,” its genuine causal power is discounted—as though a thing without thick agency could not have large effects. But causal power does not require thick agency. A system can be no one and still move markets, allocate resources, filter information for millions, select targets, or cascade a failure across infrastructure. The thinness of its agency in the philosophical sense says nothing about the thickness of its agency in the causal sense. To reason “it has no real agency, therefore it is not the kind of thing we must govern as a powerful actor” is to let the true metaphysical

modesty of the system's inner life excuse a false modesty about its outer effects.

Research Question

The two errors share a root and admit a single corrective: hold the causal question and the mental question rigidly apart, and govern the first without waiting on the second. What a system *does*—the scope of its actions, the reversibility of its effects, the domains it may act in, the thresholds at which a human must be in or on the loop—is specifiable, measurable, and governable now, on the basis of the system's demonstrated causal powers, entirely independently of whether anything is home. The mistake, in both directions, is to let the mental question bear on the causal one: to grant power because we half-believe in the person (over-attribution), or to withhold governance because we disbelieve in the person (under-attribution). The person is the wrong variable. The relevant variable is the power, and the power is present and measurable regardless of how the metaphysics resolves. The research and policy task is to build the governance of delegated causal agency on the causal facts alone—so that it is neither inflated by the illusion of a subject nor deflated by the correct observation that there is none.

4 What the Governance Literature Already Knows

The constructive response does not have to be invented; it has to be freed from the conflation. The most developed proposal is MEANINGFUL HUMAN CONTROL, which requires that an autonomous system satisfy two conditions: TRACKING—the system responds to the relevant moral reasons of the humans it serves—and TRACING—its behavior can always be traced back to at least one human who understands its capabilities and grasps the moral weight of deploying it (Santoni de Sio and Hoven 2018). Notice what these conditions do. Neither asks whether the system is a genuine agent. Both fix accountability on the human side by design, treating the system throughout as a conduit of delegated power whose actions must remain answerable to a person who could foresee and owns the stakes. The framework is, in effect, an engineering discipline for keeping power and answerability together across the joint where autonomy threatens to separate them.

That this account descends from the philosophy of moral responsibility is not incidental. Its notion of control is guidance control, drawn from the free-will debate's analysis of what makes an agent responsible for what it does (Fischer and Ravizza 1998)—the very tradition that supplies the thick sense of agency the earlier movement distinguished. The governance literature thus quietly confirms the disentangling: it does not try to make the machine responsible, because it knows the machine is not the kind of thing that can be. It keeps the responsibility where the thick agency actually is, among the humans, and asks how to design the system so that the human's responsibility remains real rather than nominal despite the machine's autonomy.

The convergent counsel

Independent lines arrive at the same posture. Accountability, on a leading analysis, decomposes into *answerability*—the duty of a named someone to account for the outcome—which cannot attach to a non-person and so must be engineered onto the human chain (Novelli et al. 2024). Recent work argues that fully autonomous agents should not be developed, because safety, privacy, and security risks rise as the human cedes control, and that bounded autonomy with retained human oversight should be the default (Mitchell et al. 2025). The catalogue of harms from increasingly agentic systems documents the concrete failure modes that emerge as delegation deepens (Chan et al. 2023). And the international assessment of AI safety defines control itself as the ability to oversee a system and adjust or halt its behavior when it acts in unwanted ways—a purely causal, purely operational criterion that makes no reference to the system’s inner life (Bengio et al. 2025; Bostrom 2014). Every one of these keeps the causal question in front and sets the mental question aside. That is the disentangling, applied.

5 The Three-Agent Architecture, Read Through the Axes

The separation defended here has a natural home in an architecture the author has developed elsewhere, and the fit is worth drawing because it shows the disentangling doing structural work. The Dual Horizon Framework distributes cognition across three agents held at distinct levels: artificial systems, institutions, and human judgment, with human judgment retained at the apex (Moritz 2026c). Read through the three axes, the architecture is precisely a refusal of the agency conflation. It assigns the artificial layer high competence and bounded, instrumented agency; it never assigns it thick agency or the answerability that would follow; and it keeps the ownership of ends—the standing to decide what the powers are *for*—at the human apex, where thick agency actually resides. The framework is, in the vocabulary of this movement, a design for granting competence and delegated causal power to systems that are not persons while keeping personhood, and the answerability that attends it, non-negotiably human.

This is the constructive inverse of the warning. The danger the movement identifies is power separated from personhood; the architecture’s response is to re-tether them deliberately, across the joint that autonomy strains, by placing the thick agent—the human—in the position from which the delegated powers are authorized, bounded, and, when necessary, halted. The metaphysics of machine mind can remain open indefinitely without disturbing this design, because the design never depended on the machine being a person. It depended only on the machine having power, which it does, and on someone answerable standing above that power, which the architecture requires.

6 Objections

“The gap is desirable.” It has been argued that responsibility gaps can be welcome, relieving humans of the burden of tragic choices no one could make well (Danaher 2016). Perhaps in narrow cases. But the movement’s point survives: whatever comfort a gap provides, it is purchased by routing consequential power through a non-bearer of responsibility, and the comfort is an argument for accepting diminished accountability, not evidence that accountability was preserved. One may sometimes choose the gap. One should not mistake it for its own absence.

“Corporations already show that non-persons bear responsibility.” The law assigns

responsibility to corporations, which are not conscious agents; why not to autonomous systems? Because corporate responsibility is a legal construct that ultimately decomposes onto human officers, shareholders, and duties—the corporation is a device for organizing human answerability, not for replacing it. To treat an autonomous system as an analogous bearer is to propose a genuinely new thing: an answerability that decomposes onto no one. That may be buildable as a legal fiction, but the fiction must be defended, not assumed—and the disentangling shows why the assumption is tempting: the homonym makes the system look like a locus of agency it does not possess.

”This waits on the metaphysics after all—if the systems became genuine agents, the analysis changes.” If a system ever acquired thick agency—genuine ownership of ends, the standing to answer—then it would become a candidate bearer of responsibility, and the analysis would indeed change. But this concedes the movement’s structure rather than resisting it. Until that transition occurs, if it ever does, the power is here and the person is not, and the governance of the power cannot be held hostage to the possible future arrival of the person. The movement that follows counsels precaution about a subject that might quietly appear. This movement counsels the opposite discipline about agency: govern the power that has unquestionably appeared, and do not let the unresolved question of the person delay it.

7 Conclusion

The felt and the owned may or may not ever arrive. The power has already arrived. The practical questions raised by the power neither require nor may await an answer to the questions raised by the person—and the conflation of agency with mind is dangerous precisely because it makes us wait, or makes us relax, when we should be building the accountability that delegated power demands.

The three conflations are not equally urgent, and this movement has argued that the agency conflation is the one that cannot wait. The intelligence conflation concerns what we might fail to build—a mind, by accident, while aiming at competence. The mind conflation concerns what we might fail to notice—a subject, present but unattested, owed a moral standing we withhold. Both are real; both can, in a sense, wait, because their objects are either not yet actual or not yet detectable. The agency conflation concerns what we have *already built and already deployed*: systems that act, with real and growing effect, and that are not the kind of thing that can answer for what they do. Its object is not hypothetical and not hidden. It is operating now.

The disentangling of the axes is what makes the situation legible. Once agency is seen as a third property, independent of competence and of mind, the practical task separates cleanly from the philosophical one. Whether these systems will ever be minds is a question that may remain open for a long time, and the census that anchors this journal exists to keep it honestly open (Moritz 2026a). But the governance of what they *do* does not depend on that question and must not be made to wait for it. We have granted power without personhood, and the standing to answer for the exercise of power belongs, still and only, to persons. The work is to keep it there—to design the delegation so that the answerability which cannot rest on the machine rests, visibly and bindingly, on the humans above it.

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

- Bengio, Y. et al. (2025). “International AI Safety Report”. In: *Department for Science, Innovation and Technology (UK)*. arXiv:2501.17805. doi: [10.48550/arXiv.2501.17805](https://doi.org/10.48550/arXiv.2501.17805).
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Chan, A., R. Salganik, A. Markelius, C. Pang, et al. (2023). “Harms from Increasingly Agentic Algorithmic Systems”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23)*, pp. 651–666. doi: [10.1145/3593013.3594033](https://doi.org/10.1145/3593013.3594033).
- Danaher, J. (2016). “Robots, Law and the Retribution Gap”. In: *Ethics and Information Technology* 18.4, pp. 299–309. doi: [10.1007/s10676-016-9403-3](https://doi.org/10.1007/s10676-016-9403-3).
- Fischer, J. M. and M. Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- 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](https://doi.org/10.1007/s10676-004-3422-1).
- Mitchell, M., A. Ghosh, A. S. Luccioni, and G. Pistilli (2025). “Fully Autonomous AI Agents Should Not Be Developed”. In: *arXiv preprint arXiv:2502.02649*. doi: [10.48550/arXiv.2502.02649](https://doi.org/10.48550/arXiv.2502.02649).
- 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](https://doi.org/10.2139/ssrn.6503099).
- (2026d). “The Third Axis: Agency, Competence, and Mind as Three Independent Things”. In: *Philadelphia Journal of Artificial and Digital Minds* 1, pp. 9–18.
- Novelli, C., M. Taddeo, and L. Floridi (2024). “Accountability in Artificial Intelligence: What It Is and How It Works”. In: *AI & Society* 39.4, pp. 1871–1882. doi: [10.1007/s00146-023-01635-y](https://doi.org/10.1007/s00146-023-01635-y).
- 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](https://doi.org/10.3389/frobt.2018.00015).
- Santoni de Sio, F. and G. Mecacci (2021). “Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them”. In: *Philosophy & Technology* 34.4, pp. 1057–1084. doi: [10.1007/s13347-021-00450-x](https://doi.org/10.1007/s13347-021-00450-x).
- Sparrow, R. (2007). “Killer Robots”. In: *Journal of Applied Philosophy* 24.1, pp. 62–77. doi: [10.1111/j.1468-5930.2007.00346.x](https://doi.org/10.1111/j.1468-5930.2007.00346.x).