PHILADELPHIA JOURNAL OF ARTIFICIAL AND DIGITAL MINDS

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

IS THE PAPERCLIP MAXIMIZER A FALSE LEAD, OR A REAL DANGER IN PLAIN CLOTHES?

Red Herrings or Ponies in Mufti

Elan Moritz
Eagles Perch Press • Philadelphia, PA

July 18, 2026

Keywords: paperclip maximizer; existential risk; AI alignment; present harms; agency; AI governance; value alignment; longtermism

About this issue

This is the fifth and final movement of the inaugural issue of the *Philadelphia Journal of Artificial and Digital Minds*, and a change of register. The first four movements built the framework—separating *competence*, *agency*, and *mind* into three independent axes, carrying the separation into the governance of delegated power, and assigning each axis its proper normative posture. This closing movement takes the completed instrument into the open and uses it on a live, heated public dispute: whether the "paperclip maximizer" and the existential-risk discourse it anchors are a distraction from AI's present harms, or a literal warning of catastrophe. It argues that both readings misidentify the danger, and that the three-axis analysis dissolves the confusion—showing the supposed opponents to be studying the same problem at different scales. Where the earlier movements construct, this one applies. Read in order, the five are one argument.

Abstract

The paperclip maximizer and its kin are increasingly dismissed, from one side, as *red herrings*: speculative distractions that divert attention and resources from the real, present harms of artificial intelligence. From the other side they are defended as literal warnings of an existential threat we must forestall. This essay argues that both readings misidentify the animal, and it borrows a construction from a companion work to say how: the maximizer is not a red herring but a *pony in mufti*—a genuine hazard traveling in plain clothes, real but disguised, present but dressed as something it is not. The distinction is not merely rhetorical. A red herring is a false lead with no quarry behind it; a pony in mufti is a real thing whose costume hides both what it is and that it can be handled. Drawing on this issue's three-axis analysis, the essay contends that the maximizer disguises a real danger *twice over*: it dresses a genuine hazard (unbounded, ungoverned agency) as a fake-looking one (a science-fiction goal monster), so that serious critics dismiss the cartoon and lose the real risk with it; and it dresses a *tractable* problem (bounding scope, reversibility, and authority—ordinary governance) as an *intractable* one (loading provably correct cosmic values before capability arrives). The red-herring reading and the literal-threat reading are two failures to see through the same mufti: one concludes there is no animal, the other panics at a dragon, and both miss that the animal is a pony one could saddle. The essay steelmans the red-herring position at its strongest—including the empirical finding that existential-risk narratives do not, in fact, measurably crowd out concern for present harms—and concludes that the honest verdict is neither dismissal nor alarm but *undressing*: strip the maximizer of its costume, and what remains is a real, mundane, governable problem that both camps, for opposite reasons, have been looking straight past.

1 Two Ways to Mistake an Animal

A red herring is a false lead with nothing behind it. A pony in mufti is a real thing in plain clothes—present, but disguised. The question about the paperclip maximizer is not whether it is frightening but which of these it is; and the answer, that it is a pony in mufti, is missed equally by those who call it a distraction and those who call it a warning.

The paperclip maximizer has become a battleground. Nick Bostrom's superintelligence that converts the reachable universe into paperclips, offered as a demonstration that supreme capability need not bring benevolent ends, has passed from philosophy into public argument, and there it has split its audience (Bostrom 2012, 2014). One camp treats it, and the existential-risk discourse it anchors, as a RED HERRING: a lurid, speculative distraction that pulls attention, funding, and regulatory will away from the real and documented harms artificial intelligence is already doing—bias, labor displacement, surveillance, the concentration of power (Bender et al. 2021; Gebru and Torres 2024). The other camp treats it as a literal forewarning: a sketch, cartoonish only on its surface, of a genuine trajectory toward an uncontrollable optimizer that we must solve the problem of value-alignment to survive (Russell 2019).

This essay proposes that the debate is malformed because both sides have misidentified the animal, and it borrows a construction from a companion work to name the error precisely. In *Wagers in Mufti*, a wager is shown to travel "in mufti"—in plain clothes, disguised as something other than the bet it is, so that people take it without recognizing its structure (Moritz 2026e). The maximizer, this essay argues, is not a red herring but a PONY IN MUFTI: a real animal in a costume. The two figures are precise opposites. A red herring is a *false* lead: follow it and there is nothing at the end, no quarry, only misdirection. A pony in mufti is a *true* thing disguised: there is an animal, and a useful one, but its plain clothes hide both what it is and—crucially—that it can be handled. To call the maximizer a red herring is to say the trail is empty. To see it as a pony in mufti is to say the trail leads somewhere real, but not where the costume suggests.

2 The Red-Herring Case, at Its Strongest

The red-herring reading deserves its best form, because it is partly right and its proponents are serious, and an essay that caricatured it would commit the very error it diagnoses in others.

The strongest version runs as follows. Artificial intelligence is, right now, producing measurable harms: discriminatory decisions, the exploitation of labeled-data labor, mass surveillance, the displacement of workers, and the concentration of power in a few firms (Bender et al. 2021). The existential-risk narrative, by contrast, rests on a speculative and unrealized capability—superintelligence—for which there is no demonstrated path, and it is entangled with an ideology, sometimes called longtermism, that weighs vast hypothetical future populations against present suffering and finds the present wanting (Gebru and Torres 2024). The worry is not only that this is unscientific but that it is *expensive*: attention and regulatory capacity are finite, and every hour spent war-gaming a paperclip apocalypse is an hour not spent on the harms already unfolding. On this reading the maximizer is worse than false; it is a *functional* distraction, whatever its authors intend, drawing the scarce resource of concern toward a phantom.

What the red-herring reading gets right

Three things, and they should be conceded plainly. **One:** the present harms are real, documented, and under-addressed, and no analysis of hypothetical futures should be allowed to obscure them. **Two:** the maximizer, taken *literally*—as a forecast that a superintelligent optimizer will shortly tile the universe—rests on capabilities not in evidence, and presenting it as a literal forecast is a genuine distortion. **Three:** the lurid framing has real costs, because it makes AI risk sound like science fiction and thereby licenses serious people to dismiss the whole subject, present harms included. On all three points the red-herring camp is correct, and the pony-in-mufti reading must preserve, not deny, them.

But the reading overreaches at one specific joint, and the overreach is now empirically testable rather than merely arguable. The red-herring case depends on a DISTRACTION HYPOTHESIS: that attention to existential risk *crowds out* attention to present harms, the two competing for a fixed pool of concern. This is an empirical claim, and it has been tested. In a large preregistered study, exposure to existential-risk framings raised concern about catastrophic risk *without* diminishing concern for immediate harms; the pool of concern did not behave as fixed, and the expected crowding-out did not appear (Hoes and Gilardi 2025). The finding does not vindicate the maximizer—it says nothing about whether the scenario is coherent—but it removes the load-bearing premise of the strong red-herring claim. If attending to the one does not measurably subtract from attending to the other, then the maximizer's sin cannot be that it *steals* concern, and the objection must retreat from "it is a distraction" to the weaker "it is mis-stated"—which is precisely the pony-in-mufti claim.

3 Why It Is Not a Red Herring: There Is an Animal

The decisive reason the maximizer is not a red herring is that an earlier movement of this issue has already found the quarry at the end of the trail—not where the costume points, but really there (Moritz 2026d).

That analysis separated three properties fused in talk of "AI"—COMPETENCE (goal-achievement capacity), AGENCY (the capacity to act, thin and thick), and MIND (experience and owned ends)—and the maximizer can be run against them (Legg and Hutter 2007; Moritz 2026d). Two findings bear directly here. First, the maximizer's menace does not lie on the axis its telling emphasizes. The story is about a *goal*—paperclips, the wrong values—but strip the system to each axis alone and the danger localizes elsewhere: competence alone is inert (an oracle that knows how to tile the universe but acts on nothing converts nothing); a wanting-but-powerless mind is harmless; everything fearful is on the *agency* axis—the unbounded, irreversible, unauthorized *action*. Give the identical paperclip objective to a system whose scope of action is bounded and reversible, and the catastrophe does not occur, with the goal unchanged. The danger was never the goal. It was the ungoverned agency.

The quarry, named

There is a real hazard at the end of the maximizer's trail, and it is this: *highly capable systems granted unbounded, irreversible, unauthorized scope of action*. That hazard does not require superintelligence, does not require consciousness, does not require the system to "want" anything in any rich sense. It requires only capability coupled to unchecked agency—which is a present and growing condition, not a speculative future one. Legal scholarship has reached the same place by another road, observing that existential-scale risks need not await superintelligence, because optimizing systems operating *within current capabilities* can produce catastrophic outcomes while remaining nominally compliant (Gropper 2024). The animal is real. It is simply not the animal the costume depicts.

This is why the red-herring verdict fails. A red herring leads nowhere; the maximizer leads to a genuine and specifiable danger, one that the governance literature already names and addresses under the headings of meaningful human control, corrigibility, scope limitation, and the refusal of full autonomy (Chan et al. 2023; Matthias 2004; Mitchell et al. 2025; Santoni de Sio and Hoven 2018). The trail is not empty. What is wrong is not that there is no quarry but that the quarry has been dressed as something else.

4 Why It Is Not the Literal Threat Either: The Costume

If the red-herring camp errs by declaring the trail empty, the literalist camp errs by mistaking the costume for the creature—by treating the science-fiction surface of the maximizer as its true form and preparing to fight a dragon.

The double disguise

The maximizer is a pony in mufti in *two* directions, and this is the essay's central observation. **Disguise one—the real dressed as the fake.** It presents a genuine hazard (unbounded agency) in the costume of a science-fiction scenario (a goal-monster tiling the cosmos), so that serious critics, rightly rejecting the science fiction, discard the real hazard along with it. The pony is dressed as a clown, and so is laughed out of the room. **Disguise two—the tractable dressed as the intractable.** It presents a governable problem (bound the scope, require reversibility, retain authority) in the costume of an unsolved cosmic one (specify perfect values, forever, in advance), so that the problem looks too hard to attempt. The pony is dressed as a dragon, and so is fled from rather than saddled. The red-herring reader sees through disguise one and concludes there is no animal. The literalist reader is taken in by disguise two and prepares for a monster. Both miss that the costume hides a pony—real, and rideable.

Taken literally, the maximizer directs us to solve VALUE ALIGNMENT: to load provably correct final goals into a system before it becomes capable enough to resist, because a sufficiently intelligent optimizer will pursue whatever end it was given to catastrophic completion. But the three-axis analysis showed this target to be misplaced for the very system the maximizer describes. If the maximizer is a *mindless* optimizer—the version that makes it plausible as a near-term worry—then it has no thick agency, no ownership of ends, and therefore *no seat into which values could be loaded*; "align its goal" addresses an axis the system does not occupy (Moritz 2026d). And if instead it is a genuine agent that owns its end, then it can reflect on that end, and the guaranteed, unreflective fixity that made it terrifying is gone. Either way,

provably-correct value-loading is aimed at the wrong place: at the goal, when the danger is in the scope of action, and at a seat of ends the plausible version does not have.

5 The Cost of the Costume

Misdirection is not free, and the maximizer's costume has exacted a measurable price from both camps and from the field between them.

From the alignment program, the cost has been a decade of effort trained substantially on *value specification*—on how to load the right goal—for systems whose danger, the axis analysis suggests, lies in the *agency envelope* rather than the goal (Russell 2019). Effort spent proving values correct is effort not spent bounding scope, building corrigibility, and instrumenting reversibility —the interventions that act where the danger actually is (Mitchell et al. 2025; Santoni de Sio and Hoven 2018). The costume did not merely mislabel the problem; it redirected the labor.

From the critics, the cost has been the discarding of a real concern along with the cartoon. Because the hazard arrived dressed as science fiction, those most attentive to AI's present harms have often rejected the entire register of "AI risk" as ideology, and with it the agency-governance problem that is in fact continuous with their own concerns (Bender et al. 2021; Gebru and Torres 2024). This is the deeper irony the pony-in-mufti reading exposes: the ungoverned-agency hazard is *not* a rival to the present-harms agenda but an extension of it. Concentration of power, unaccountable automated decision-making, and the deployment of consequential systems without meaningful human control are agency-governance problems whether the system is a credit-scoring model today or a broadly capable agent tomorrow. The maximizer's costume made a natural ally look like an enemy.

Research Question

The reframing suggests that the productive question is neither "is AI an existential risk?" nor "is existential-risk talk a distraction?" but *"what is the governance of delegated agency, across the whole capability range, from today's deployed systems to tomorrow's more autonomous ones?"* Posed this way, the supposed opposition between present-harms researchers and existential-risk researchers largely dissolves: both are, or should be, studying the same axis—the scope, reversibility, authorization, and oversight of consequential action by capable systems—at different points along it. The empirical finding that existential framings do not crowd out present-harm concern (Hoes and Gilardi 2025) is consonant with this: the two are not competitors for a fixed pool of attention because they are, correctly understood, the same problem at different scales. The open research task is to build a single governance framework indexed by *degree of delegated agency* rather than two rival discourses indexed by *timeline of harm*—a framework in which the credit-scoring model and the autonomous agent are near and far points on one axis, governed by one set of principles about what capable systems may be permitted to do without a human answerable in the loop (Moritz 2026b,c). Whether such a unified framework can be built, and whether it can hold the political coalition that the maximizer's costume has kept divided, is the question this reframing opens—and a fitting one on which to close the issue that built the instrument for posing it.

6 Objections

"The maximizer was always an existence proof, not a forecast, so calling it a costume misreads its purpose." Its defenders are right that Bostrom offered it to refute a specific assumption—that sufficient intelligence brings benevolent ends—and as that refutation it succeeded (Bostrom 2012). But an existence proof that becomes the field's central image does work beyond its original charge, and the work it now does is misdirection: it trains attention on the goal axis and the value-loading problem, in public argument and in research priorities alike. The pony-in-mufti reading does not deny the maximizer's original success as an intuition pump (Dennett 2013); it observes that the pump, having killed the "intelligence implies wisdom" assumption, is now asked to localize the danger, and localizes it wrongly. One can honor the existence proof and still undress the costume it has since become.

"This smuggles in the claim that superintelligence is not a real risk, which the analysis has not established." It does not, and the distinction matters. The pony-in-mufti reading is silent on whether superintelligence will arrive and neutral on its probability. Its claim is narrower and more robust: *whatever* the capability level, the danger localizes on the agency axis, so that the governance of scope and authority is the operative variable across the entire range. If superintelligence arrives, the agency-governance problem becomes harder and more urgent, not different in kind—the question remains what a capable system may be permitted to do without an answerable human in the loop. The reading neither inflates nor deflates the probability of superintelligence; it relocates the danger in a way that holds regardless.

"If present harms and existential risk are really one problem, the distinction collapses and the critics were right to focus on the present." The problems are continuous on the agency axis but not identical in urgency or tractability, and the continuity is an argument for *joining* the agendas, not for abandoning either. The present-harms researchers have the better claim on *what is documented now*; the risk researchers have the better claim on *how the same hazard scales with capability*. The pony-in-mufti reading asks each to recognize the other as working the same axis at a different point—which is the opposite of telling either to stand down. The costume made them look like rivals; undressing it makes them colleagues.

7 Conclusion

The paperclip maximizer is neither a red herring nor a dragon. It is a pony in mufti: a real, mundane, governable hazard—unbounded agency in capable systems—traveling disguised as a science-fiction goal-monster. The costume misleads twice, dressing the real as fake and the tractable as intractable, and both camps in the debate have been deceived by one of its two disguises. The honest response is not dismissal and not alarm but undressing.

Asked whether the maximizer and its kin are red herrings, the answer this essay defends is: no, but not for the reason the literalists give. They are not red herrings because there is a real animal at the end of the trail—the hazard of highly capable systems wielding unbounded, unaccountable scope of action, a hazard that requires neither superintelligence nor consciousness and is continuous with the present harms the maximizer's critics rightly emphasize. And they are not the literal existential dragons the alarmists depict, because the danger sits on the agency axis rather than the goal axis, which means it is not a cosmic value-specification problem but an ordinary, if hard, governance problem—boundable, reversible, authorizable, the kind of thing one saddles rather than flees. The maximizer is a real thing in plain clothes, and its clothes have fooled everyone: the critics into seeing an empty distraction, the believers into seeing an unstoppable monster.

The construction that names the error comes from a companion study of wagers that travel in mufti—real bets disguised as something safer, taken by people who do not see the structure they are accepting (Moritz 2026e). The paperclip maximizer is the same phenomenon in the register of risk: a real danger disguised as something *less* tractable and *more* fantastical than it is, so that the people best equipped to address it either laugh or panic, and few undress it. To undress it is to find, under the science-fiction costume, a pony—the plain, real, rideable problem of governing what capable systems are permitted to do.

This issue began by separating the properties our talk of minds and machines fuses: competence, agency, and mind. Four movements built that separation and drew its lessons; this last one has taken it into a public quarrel and found the quarrel to be a costume—the animal both sides argue over is the same pony, and it was wearing mufti all along. That is the whole use of the instrument the issue set out to build: not to settle what artificial minds are, which remains gloriously open, but to keep us from mistaking, in the things we make, the clown for the danger and the dragon for the problem we could actually solve (Moritz 2026a,d).

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

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 610–623. DOI: 10.1145/3442188.3445922.

Bostrom, N. (2012). "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". In: *Minds and Machines* 22.2, pp. 71–85. DOI: 10.1007/s11023-012-9281-3.

– (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.

Dennett, D. C. (2013). "Intuition Pumps and Other Tools for Thinking". In: New York: W. W. Norton.

Gebru, T. and É. P. Torres (2024). "The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence". In: *First Monday* 29.4. DOI: 10.5210/fm.v29i4.13636.

Gropper, J. (2024). "The Synthetic Outlaw: Optimizing Systems and the Limits of Compliance". In: *SSRN Working Paper*. On existential-scale risk from systems operating within current capabilities.

Hoes, E. and F. Gilardi (2025). "Existential Risk Narratives about AI Do Not Distract from Its Immediate Harms". In: *Proceedings of the National Academy of Sciences* 122.16, e2419055122. DOI: 10.1073/pnas.2419055122.

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.

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.

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

– (2026b). "Precaution and Non-Deferral: Why the Same Disentangling Yields Opposite Counsels for Mind and Agency". In: *Philadelphia Journal of Artificial and Digital Minds* 1.

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

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

– (2026e). *Wagers in Mufti*. Source of the "in mufti" construction: a real structure traveling in plain clothes. Philadelphia: Eagles Perch Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

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.