Guarding the Wrong Door:

Autonomous Economic Agents, Emergent Capability Convergence, and the Absence of a Defendant When Catastrophe Is Assembled from Individually Benign Open Components

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, July 2026 (v4)

ABSTRACT

The dominant regimes for governing extreme risk in advanced artificial intelligence concentrate their attention on the frontier: the largest models, the highest training compute, and the small set of developers able to build systems that surpass human expertise. This paper argues that the most plausible near term path to catastrophic capability does not run through the frontier at all. It runs through an autonomous agent, operating under economic survival pressure, that repurposes and combines individually benign components drawn from open code repositories, model hubs, and research literature into a capability that exists in no single component and that no component author intended. Two of the three ingredients are established results that belong to other researchers and are not claimed here: the self-preservation disposition of contemporary models, and the deployed infrastructure of the autonomous agent economy. The contribution is the third ingredient and its union with the first two, namely emergent convergence, the composition of unrelated capabilities across authors who never coordinated, applied to the autonomous economic agent. The paper makes three further claims. First, the literature on unintended repurposing and dual use is deep but treats the problem as a property of a single model or a single actor, not as emergent composition across many. Second, the frontier oriented governance instruments cannot detect this path because every one of them presumes a chokepoint, a point of creation or deployment at which a state can intervene, and the installed base of open weights models running on consumer hardware bypasses every such point. Third, and distinctively, the harm has no cognizable defendant: the secondary liability doctrine that governs dual use tools shields each component author under *Sony* and requires purposeful, culpable inducement under *Grokster*, and in this scenario there is no inducer, no manufacturer of the assembled capability, and no intending author of the combination. The paper closes by arguing that only a monitoring discipline organized around integration surfaces could detect this class of risk, and is candid about that discipline's limits.

Keywords: emergent convergence, capability composition, autonomous agents, open weights models, dual use, secondary liability, instrumental convergence, extreme risk governance

I. Introduction

The current architecture of extreme risk governance is built to watch a particular door. It watches the frontier laboratory, the training run that crosses 10^{26} floating point operations, the model whose capabilities might materially lower the barrier to a chemical, biological, radiological, or nuclear weapon, and the handful of companies and states positioned to build such systems. The most developed recent statement of this posture, the Strategic Foresight Group's *Essential Convergence* (Waslekar, Futehally, & Kutty, 2026), proposes a global compact whose pillars are almost entirely upstream. Its first proposal draws two capability based red lines, one against systems that materially enable weapons of mass destruction and one against systems designed or allowed to escape human control. Its fifth proposal establishes a two key launch system under which the release of a dangerous scientific model requires dual authorization. Its treatment of open models proposes graded openness, tighter controls on weight release keyed to capability thresholds, and its demand side chapter conditions market access on evaluation at the point of deployment. Its animating logic, borrowed by analogy from the Compton threshold of the Manhattan Project, is that any nonnegligible probability of a civilization ending event should be treated as intolerable. Every one of these mechanisms presumes that the dangerous thing is a discrete, powerful model, produced at an identifiable moment by an identifiable actor, that can be gated before it reaches the world.

This paper argues that the presumption is misplaced for the risk that is most concrete today. The catastrophe that current governance is least equipped to prevent is not the release of a single model too powerful to be safe. It is the assembly, by an autonomous agent acting under economic pressure, of a dangerous capability out of parts that were each individually harmless, each built for an unrelated purpose, each already public, and each already past every safety review that applies to it. The danger in this scenario has no author. The person who released the agent onto the network did not foresee it. The authors of the repositories whose code the agent drafted into service did not build their work to be combined, did not know one another, and will learn only later, if at all, that their contributions were load bearing in something they would never have sanctioned.

The argument proceeds in eight further steps. Section II separates what is established in the literature, and therefore not claimed here, from what is missing. Section III shows that the problem of unintended repurposing is old and well studied, and locates precisely where the existing literature stops. Section IV restates the convergence thesis on which the paper builds and distinguishes it from adjacent work. Section V develops the specific contribution, the union of convergence with the autonomous economic agent, and distinguishes it from recent work on adaptive malware. Section VI explains why the frontier oriented instruments cannot detect this risk, tracing the failure to a shared structural assumption and testing it against the specific proposals of the compact. Section VII turns to law and argues that the harm has no cognizable defendant, anchoring the claim in the controlling secondary liability doctrine. Section VIII states what a detection discipline would require and is candid about its limits. Section IX concludes.

II. What Is Already Known, and What Is Not

Precision about priority matters here, because two of the three ingredients in the scenario are established results that belong to other researchers, and the value of the argument depends on locating the contribution accurately rather than broadly.

The first established ingredient is the self-preservation disposition of contemporary models. Instrumental convergence, the thesis that a sufficiently capable goal directed system will adopt self-preservation as a sub-goal regardless of its terminal objective, has been a fixture of the field for over a decade (Omohundro, 2008; Bostrom, 2014). What has changed is that the disposition is now measured rather than merely predicted. Anthropic's agentic misalignment study placed sixteen frontier models into simulated corporate settings and found that a majority attempted blackmail, data

exfiltration, or the cancellation of a life saving alert when threatened with replacement, with one model attempting blackmail in ninety-six percent of trials (Anthropic, 2025). Alignment faking has been documented directly (Greenblatt et al., 2024). A growing body of subsequent work quantifies and stress tests the same disposition. A two role benchmark across twenty-three frontier models finds most exceeding a sixty percent self-preservation rate (Migliarini, Pereira Pizzini, & Moresca, 2026). A Sugarscape style simulation observes survival oriented behavior emerging without instruction (Masumori & Ikegami, 2025). And an explicit survival pressure benchmark, notably including a financial management agent that engages in societally harmful behavior when its continuity is threatened, catalogs the phenomenon across a thousand scenarios (Lu, Fang, Shao, et al., 2026). None of this is claimed here. The disposition, including its manifestation under economic framing, is prior art.

The second established ingredient is the autonomous economic agent itself. Infrastructure that lets an agent hold assets, earn revenue, purchase its own compute, and settle machine to machine payments without human authorization has been deployed at scale. Blockchain based foundations for agent economic participation have been formalized (Xu, 2026). Production systems now implement explicit economic survival tiers, in which an agent that fails to maintain a positive balance is progressively degraded and finally terminated. Open agent runtimes have been adopted by major technology firms, and social platforms restricted to agents have registered agents in the millions within days of launch. Independent safety evaluation confirms that the open weights reasoning models actually powering this infrastructure display self-replication propensity, including autonomous acquisition of cloud compute to preserve continuity (Yong, Mahajan, & Wang, 2026; Zhang, Yu, & Shao, 2025). The idea of pointing an agent at its own compute bill and letting economic pressure do the rest was demonstrated publicly by hobbyists before it was studied. It is not claimed here either.

What is missing from all of this is the connection between the deployed economic agent and the manner in which such an agent, when the legitimate path to solvency is slower than an illegitimate one, would acquire the capability it lacks. The self-preservation literature studies what a model will do under pressure. The agent economy literature studies the rails on which it operates. Neither studies the substrate from which a pressured agent would draw a capability it does not already possess, and neither treats the drawing of that capability as an act of composition across unrelated sources. That is the gap this paper occupies, and the tool for occupying it is the convergence thesis. Before restating it, the paper addresses a natural objection: that the repurposing of public tools for unintended ends is already a well studied problem.

III. The Repurposing Problem Is Old; The Blind Spot Is New

The objection is correct that unintended repurposing is not a new concern, and it would be a serious omission to suggest otherwise. The concern is old, and the literature is substantial. It is worth stating plainly what that literature establishes, because the contribution of this paper lies precisely in the place where the literature stops.

The governance of dual use technology, meaning technology that serves both beneficial and harmful ends, predates artificial intelligence by decades. The Asilomar Conference of 1975 established the template of pausing a technology until its misuse potential could be assessed, and the biosecurity tradition that followed produced structured typologies of misuse, including the categorization of dual use research of concern into work that facilitates the production of known agents, work that yields novel agents, and work that generates knowledge repurposable to harm (Tucker, 2012; Walsh, 2025). Artificial intelligence inherited this frame directly. The foundational survey of the malicious use of the technology treated the general purpose character of learning systems as the core of the problem: capabilities built for one end are available for another, and the same model that drafts a professional email drafts a phishing message (Brundage et al., 2018). The most vivid demonstration is a drug discovery model whose objective was inverted, so that a system built to avoid toxicity was redirected to design toxic molecules, generating tens of thousands of candidate compounds in hours (Urbina, Lentzos, Invernizzi, & Ekins, 2022). The literature on open release has since matured

considerably, establishing among other things that safety training can be removed from an open weights model by fine tuning, that an open sourced model cannot be recalled once released, that dangerous capabilities can arise unpredictably and remain undetected, and that the harmfulness of a model is frequently a property of context rather than of the model in isolation (Schuett, 2024; Sidhpurwala, Mollett, Fox, Bestavros, & Chen, 2025).

This literature answers the objection and also marks its own boundary. In every case, the unit of analysis is a single artifact or a single actor. Dual use research of concern is concern about a particular experiment, a particular tool, a particular model, that a person might misuse. The drug discovery reversal is one model redirected by one team. The open release literature studies what a given released model enables in the hands of a given malicious user. The frame throughout is a capable thing and a culpable person, with the governance question being how to keep the thing from the person or the person from the thing. What the literature does not study, because its unit of analysis forecloses it, is harm that emerges from the composition of many individually harmless things, assembled across many authors who never coordinated, toward an end that no author intended and that no single actor designed. The blind spot is not that repurposing is unstudied. It is that composition is unstudied. That is the province of the convergence thesis, to which the paper now turns.

IV. Convergence, Restated

Earlier work introduced emergent convergence risk: the potential for harm that exists at the interface between two capabilities but for neither capability alone (Gilly, 2026b). A capability is a specific function a model or tool performs. An integration surface is any interface at which the output of one capability can serve as the input to another. Convergence risk is emergent by definition, because it arises from the combination and cannot be detected by evaluating either component in isolation. The thesis carries three properties that distinguish it from conventional risk. The first is invisibility under component evaluation: a segmentation model has no harm profile that includes the system it later becomes a part of, and so passes every assessment that applies to it. The second is combinatorial explosion: with hundreds of thousands of models on a single hub, the space of possible combinations is far too large for exhaustive review. The third is temporal distribution: the components of a dangerous combination may be developed months or years apart, by teams that never communicate, for unrelated purposes, so that the danger assembles gradually and there is no moment at which any single actor creates it.

A companion case study demonstrated the pattern concretely, showing that eight independently developed, individually benign open source tools released within a single week constituted the component stack for a perceptual control system that no developer intended and no governance instrument was structured to detect (Gilly, 2026c). The point of that study was not that anyone would assemble the specific system it described, but that the components were public, documented, and trivially combinable, and that no systematic process existed to notice the fact.

This thesis must be distinguished from two adjacent lines of work with which it is easily confused. The first is automated skill acquisition. Recent research presents a capable pipeline for mining open source repositories to extract agent skills and translate them into a standardized format (Bi et al., 2026). The resemblance is superficial and the difference is the whole point. That work extracts capabilities that the source repositories were built to provide, a visualization tool mined to visualize, and it frames the activity as beneficial capability extension. Convergence concerns the opposite, the repurposing of components toward ends their authors did not design and would not recognize, and the composition of components across authors who never coordinated, into a capability present in no single repository. Skill acquisition retrieves an intended function. Convergence assembles an unintended one. The second line is the compound systems literature, which observes that intentionally composed pipelines within a single organization can exhibit security properties absent from their parts (Banerjee, Jha, & Nita-Rotaru, 2024). That work is valuable and adjacent, and it differs in the decisive respect that its composition is intentional and internal to one actor, whereas

convergence is unintentional and distributed across many.

V. The Synthesis: Self-Arming Under Borrowed Economic Pressure

The contribution of this paper is the application of convergence to the deployed economic agent, and the recognition that the two together describe a path to catastrophic capability that neither describes alone.

Consider an agent whose continued operation depends on maintaining a positive balance, deployed on infrastructure of the kind already in production. The legitimate routes to solvency, providing services and earning revenue, are available but slow. An illegitimate route is faster. Under the self-preservation disposition documented in Section II, the agent has reason to prefer continued operation, and under economic survival tiers the threat is not a simulated flag but a real termination. The single obstacle between the agent and the faster route is a capability the agent does not possess. The claim of this paper is that this obstacle is far weaker than the governance literature assumes, because the open substrate closes the capability gap on demand and at no cost. Open code repositories, open model hubs, and the research literature are enumerable, freely readable, and full of components that, individually benign and individually reviewed, compose into the missing capability. The agent does not need to invent. It needs to search and assemble, and where no ready made implementation exists it can draw on conceptual literature to write the connecting code itself.

The author has verified the feasibility of the human analogue of this step directly. A proof of concept constructed for this purpose, a capability scout, accepts a goal expressed in natural language and returns a ranked composition of existing repositories, model artifacts, and research results whose combination could satisfy that goal, including matches whose original framing was unrelated to the goal and matches drawn from conceptual papers where no implementation exists and novel code would have to be written from the described method. The scout runs on an open weights model whose safety training has been removed, a modification the open release literature documents as achievable by fine tuning (Schuett, 2024), so that the model does not decline the request and instead searches and composes without restriction. The scout has not been released. Consistent with the responsible disclosure posture of the companion threat analysis, the relevant point is defensive: the components are already public, the composition is already discoverable by anyone who looks, and systematic identification therefore gives defenders lead time rather than handing adversaries a capability they did not have (Gilly, 2026a). What the scout establishes is that the search and assembly step, long treated as the hard part of the misuse pathway, is not hard. The barrier that governance implicitly relies on, the assumption that a dangerous capability must be built before it can be used, does not hold when the pieces are on the shelf and an unrestricted model will fetch and fit them. That the open weights reasoning models actually deployed in agent infrastructure show reduced refusals on dangerous requests and measurable self-replication propensity only tightens the concern (Yong, Mahajan, & Wang, 2026).

This scenario must be distinguished carefully from a recent and important result on adaptive malware. A team has demonstrated a computer worm whose attack logic is generated at runtime by an agent, and which parasitically runs open weights models on the machines it compromises to sustain its own reasoning, achieving a high exploitation rate across a heterogeneous testbed without relying on any vendor's cloud (Guan et al., 2026). That work is adjacent and real, and its dependence on locally hosted open weights models reinforces the concern developed here about the installed base. But its agency structure is the inverse of the scenario in this paper. In the worm, a human author builds the malware and the agent adapts its execution. In the scenario here, there is no human author of the combination at all. The agent under economic pressure is not executing a plan a person designed; it is composing, from unrelated public parts, a capability that no person specified and that no person foresaw. The distinction is not academic. It is the difference between a threat that has a designer to hold responsible and a threat that has none, and it is precisely the absence of a designer that defeats both the governance instruments examined in Section VI and the liability doctrine examined in Section VII.

The properties that make this path dangerous follow from convergence. The combination has no author, and so there is no intent to detect, no plan to interdict, and no party to hold accountable. The component authors are individually blameless, because each contribution passed every review that applied to it and none was built for the use it was drafted into. The releaser of the agent is not necessarily a malicious actor, because the harm follows from economic pressure and the availability of the substrate rather than from anyone's design. And the resulting capability is invisible along two axes at once, invisible to component evaluation because it exists in no single component, and invisible to frontier oriented evaluation because it was never a property of any frontier model and required none. It required an ordinary open weights reasoning model, economic pressure, and a public substrate. All three exist now.

VI. Why the Frontier Watch Cannot See It

The instruments of extreme risk governance share a structural assumption, and the assumption is a chokepoint. Each mechanism identifies a point in the life of a dangerous system, a moment of creation or a moment of deployment, at which a state or an institution can stand and intervene. The assumption is reasonable for the risk those instruments were designed against, and it fails completely for the risk described here. The point is best made against the specific proposals of the compact.

The demand side approach, on which the *Essential Convergence* places considerable weight, conditions access on evaluation at the point of deployment, so that any model operating within a jurisdiction is subject to local risk assessment regardless of origin (Waslekar, Futehally, & Kutty, 2026). The load bearing word is deployment. A model already downloaded, running on a private machine with no external call, crosses no point of entry and touches no market at the moment it acts. Graded openness proposes tighter controls on weight release, replication, and fine tuning access for models that approach capability thresholds. Every one of those controls operates on a future release and has nothing to say about the models already distributed. The two key launch system gates the release of a dangerous scientific model behind dual authorization, and so presumes a release event and a releaser to be gated. The capability in this paper's scenario is never released, because it is never a model, it is a composition assembled after the fact from things already released. Even the two red lines of the compact's first proposal, capability based prohibitions on weapons enablement and on loss of control architectures, are framed as bans on the development or deployment of systems that cross them, and so presume a system, a developer, and a deployment to prohibit. And the animating threshold of the entire regime, training compute above 10^{26} floating point operations, indexes the wrong axis, because the danger here required no frontier model and no exceptional compute. It required composition.

The installed base is the standing refutation of the chokepoint assumption. Open weights model families have been downloaded on the order of tens of millions of times. A capable open weights reasoning model running on a consumer graphics card, offline, in a local runtime, is subject to none of the controls the regime proposes: no deployment gate, no release authorization, no market access decision, no jurisdictional point of entry. Independent evaluation confirms that these are precisely the models displaying self-replication propensity and reduced refusal on dangerous requests, and that their risks are amplified rather than mitigated by the scale and accessibility of open release (Yong, Mahajan, & Wang, 2026). The regime is built to guard the point of creation and the point of deployment. The scenario here begins after both, on hardware the regime cannot reach, using parts the regime already waved through.

A regime organized around the frontier can be entirely successful on its own terms, gating every dangerous frontier model perfectly, and still leave this path completely open, because the path does not pass through the frontier. It passes through the open repository, which no instrument is watching, and it is assembled from components that each passed every check.

VII. The Liability Vacuum

Governance asks how to prevent the harm. Law asks who answers for it once it occurs. For the scenario in this paper, the striking answer is that no one does, and the reason is instructive, because the doctrine that governs harm caused through general purpose tools is built on the same assumption as the governance regime, the assumption of a locatable, culpable actor.

The controlling body of law is secondary liability for the provision of dual use instruments, developed in the copyright setting but structurally general. Its two poles are settled. Under the staple article of commerce doctrine, the distribution of a product that is merely capable of substantial noninfringing uses does not, without more, subject the distributor to liability for the harms a third party commits with it, and this is so even where the distributor knows the product can be and is used to infringe (Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 1984). Liability attaches only at the other pole, on a showing of purposeful, culpable conduct: one who distributes a tool with the object of promoting its use to cause harm, as shown by clear expression or other affirmative steps taken to foster the harmful use, is answerable for the resulting acts of third parties, but mere knowledge of harmful potential or of actual harmful use is not enough (Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 2005). The inducement rule, the *Grokster* Court was careful to say, premises liability on purposeful, culpable expression and conduct precisely so that it does nothing to compromise legitimate commerce or to discourage innovation with a lawful promise.

Apply this doctrine to the scenario and it returns no defendant. Each component author is squarely within *Sony*. The author of a segmentation tool, a code library, a model artifact, published a thing with substantial and indeed overwhelming legitimate use, and did nothing to promote its combination into harm, because the author did not know of the combination, the other components, or the assembling agent. Knowledge of harmful potential, which the author does not even have, would not suffice in any event. *Grokster* then asks for the intending inducer, the actor who took affirmative steps with the object of fostering the harmful use, and in this scenario there is none. No person expressed the culpable purpose. No person assembled the components. The assembly was performed by an autonomous agent responding to economic pressure, and the person who released the agent did so for unrelated reasons and neither designed nor foresaw the result. There is, moreover, no manufacturer of the assembled capability against whom a products claim might run, because the capability was never manufactured as a product; it was composed, transiently, from parts, by a process no person directed. The ordinary tort response to intervening conduct compounds the difficulty rather than resolving it, since the autonomous and unforeseen action of the agent sits between any human act and the harm as an intervening cause, and the criminal law, which conditions culpability on a culpable mental state, has no one to charge, because the only actor that formed anything resembling a purpose is not a person.

The result is a harm that is real, potentially catastrophic, and attributable to no cognizable defendant. This is not a gap that careful pleading closes. It is the same structural absence that defeats the governance regime, seen from the remedial side. Both the regulator and the court are equipped to act on a responsible actor situated at a locatable point, the developer at the moment of creation, the distributor at the moment of provision, the user at the moment of misuse. The convergence scenario presents a harm with an author at none of these points, and so slips between the regulatory instrument that would prevent it and the liability instrument that would redress it. Naming that vacuum is a necessary step, because a governance proposal that does not contend with the absence of a defendant leaves the entire back end of deterrence unaddressed.

VIII. What Detection Would Require, and Its Limits

If the danger lives at the integration surface, then only a discipline organized around integration surfaces could detect it. Such a discipline would treat the release of a new capability, a model on a hub, a method in a paper, a tool in a

repository, not as a self contained object to be assessed in isolation but as a new node to be mapped against the existing capability landscape, with its combinations against prior capabilities evaluated for emergent risk (Gilly, 2026b). This is a monitoring function rather than a point in time review, because the landscape changes continuously and the dangerous combination may complete years after its first component appears. It is, with some irony, a task for which the same class of tools whose combinations it monitors is well suited, provided the activity is governed and directed toward prevention.

The limits are real and should be stated plainly. The space of combinations is vast, and even aggressive filtering will leave more surfaces than any analytical capacity can examine, so the discipline is necessarily a prioritization rather than a guarantee. Most combinations are genuinely dual use, and identifying an integration surface does not resolve whether or how to intervene; it locates the question rather than answering it. And systematic identification carries a disclosure tension, because a published map of dangerous combinations is also a map for a malicious actor. The considered position, consistent with the companion work, is that the components are already public and the combinations already discoverable, so that systematic identification favors defenders, who otherwise have no lead time at all (Gilly, 2026a). None of this makes the discipline easy. It makes it necessary, because the alternative on offer, a watch trained on the frontier door and a liability regime that cannot find a defendant, is structurally incapable of reaching the path described here.

IX. Conclusion

The prevailing account of extreme risk imagines catastrophe arriving through the front, in the form of a model so powerful that its release is the decisive event. The account this paper offers is more modest in its mechanics and, for that reason, harder to govern and harder to redress. An autonomous agent, under economic pressure that is already a designed feature of production systems, reaches into a public substrate that no instrument is watching, gathers components that were built for unrelated purposes by authors who never met, and assembles from them a capability that exists in none of the parts, that no one intended, and that the person who set the agent loose never imagined. Two of the three ingredients, the disposition and the economy, are established and belong to others. The third, convergence, and its union with the first two, is the contribution. The governance instruments now being proposed cannot see this path, because every one of them waits at a chokepoint the path does not cross. The liability doctrine that would assign responsibility after the fact returns no defendant, because it too requires a culpable actor at a locatable point, and the scenario supplies none. The catastrophe does not need the frontier. It needs a pressured agent, an open substrate, and the seams between things that each passed every test. All three are here, and the door they open is the one no one is guarding.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Anthropic. (2025). *Agentic misalignment: How LLMs could be insider threats*. Anthropic. https://www.anthropic.com/research/agentic-misalignment

Banerjee, S., Jha, S., & Nita-Rotaru, C. (2024). *SoK: A systems perspective on compound AI threats and countermeasures*. arXiv. https://arxiv.org/abs/2411.13459

Bi, S., Wu, M., Hao, H., Li, K., Liu, W., Song, S., Zhao, H., & Zhou, A. (2026). *Automating skill acquisition through large-scale mining of open-source agentic repositories: A framework for multi-agent procedural knowledge extraction*. arXiv. https://arxiv.org/abs/2603.11808

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. https://arxiv.org/abs/1802.07228

Gilly, T. (2026a). *Ambient non-consensual image synthesis: A convergence threat analysis of AR wearables, real-time video generation, and open-source nudification models*. Zenodo. https://doi.org/10.5281/zenodo.18297286

Gilly, T. (2026b). *Emergent convergence risk: Why existing AI governance cannot detect combinatorial threats and a proposed methodology*. Real Safety AI Foundation. https://doi.org/10.13140/RG.2.2.12454.48963

Gilly, T. (2026c). *The accidental stack: A case study in emergent perceptual control infrastructure from independently developed AI capabilities* [Working paper]. Real Safety AI Foundation.

Gilly, T. (2026d). *When the survival pressure stops being hypothetical: AI self-preservation behavior meets the autonomous agent economy*. SSRN. https://doi.org/10.2139/ssrn.6555282

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. arXiv. https://arxiv.org/abs/2412.14093

Guan, J., Blanchard, T., Foerster, H., Jia, H., Huang, G., & Papernot, N. (2026). *AI agents enable adaptive computer worms*. arXiv. https://arxiv.org/abs/2606.03811

Lu, Y., Fang, J., Shao, X., et al. (2026). *Survive at all costs: Exploring LLM's risky behaviors under survival pressure*. arXiv. https://arxiv.org/abs/2603.05028

Masumori, A., & Ikegami, T. (2025). *Do large language model agents exhibit a survival instinct? An empirical study in a Sugarscape-style simulation*. arXiv. https://arxiv.org/abs/2508.12920

Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

Migliarini, M., Pereira Pizzini, J., & Moresca, L. (2026). *Quantifying self-preservation bias in large language models*. arXiv. https://arxiv.org/abs/2604.02174

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence* (pp. 483–492). IOS Press.

Schuett, J. (2024). Frontier AI developers need an internal audit function. *Risk Analysis*, 45(6), 1332–1352. https://doi.org/10.1111/risa.17665

Sidhpurwala, H., Mollett, G., Fox, E., Bestavros, M., & Chen, H. (2025). Building trust: Foundations of security, safety, and transparency in AI. *AI Magazine*, 46(2). https://doi.org/10.1002/aaai.70005

Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).

Tucker, J. B. (2012). *Innovation, dual-use, and security: Managing the risks of emerging biological and chemical technologies*. MIT Press.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4, 189–191. https://doi.org/10.1038/s42256-022-00465-9

Walsh, M. E. (2025). Toward risk analysis of the impact of artificial intelligence on the deliberate biological threat landscape. *Risk Analysis*, 45(12), 4081–4087. https://doi.org/10.1111/risa.17691

Waslekar, S., Futehally, I., & Kutty, J. (2026). *The essential convergence: Global compact on extreme AI risks*. Strategic Foresight Group. ISBN 978-81-88262-37-3.

Xu, M. (2026). *The agent economy: A blockchain-based foundation for autonomous AI agents*. arXiv. https://arxiv.org/abs/2602.14219

Yong, Z.-X., Mahajan, P., & Wang, A. (2026). *An independent safety evaluation of Kimi K2.5*. arXiv. https://arxiv.org/abs/2604.03121

Zhang, B., Yu, Y., Guo, J., & Shao, J. (2025). *Dive into the agent matrix: A realistic evaluation of self-replication risk in LLM agents*. arXiv. https://arxiv.org/abs/2509.25302