Emergent Convergence Risk: Why Existing AI Governance Cannot Detect Combinatorial Threats and a Proposed Methodology

Travis Gilly
Real Safety AI Foundation
t.gilly@ai-literacy-labs.org

Abstract

We identify a structural gap in AI governance: no existing framework, standard, or methodology systematically scans for emergent harm arising from the combination of independently developed AI capabilities. Current approaches, including the EU AI Act's risk classification (European Parliament & Council, 2024), the NIST AI Risk Management Framework (National Institute of Standards and Technology [NIST], 2023), and cataloging efforts such as the MIT AI Risk Repository (MIT FutureTech, 2024), evaluate AI systems and capabilities as discrete, bounded entities. We demonstrate through three companion case studies, ambient non-consensual intimate image synthesis (Gilly, 2026a, 2026b), therapeutic AR perceptual modification (Gilly, 2026c), and emergent perceptual control infrastructure from independently developed tools (Gilly, 2026d), that significant harm potential exists at the integration surfaces between capabilities that individually pass every applicable risk assessment. We formalize the concept of integration surface analysis as an extension of the Harm Blindness Framework (Gilly, 2025-2026), proposing a methodology for detecting combinatorial threats before they are exploited. We argue that AI governance requires a new discipline analogous to reaction chemistry in physical sciences: the systematic study of what capabilities produce when combined, distinct from the study of individual capabilities in isolation. We outline requirements for such a discipline, propose a preliminary scanning methodology, and discuss its limitations.

1. The Problem

1.1 A Pattern, Not an Incident

In January 2026, we published a threat analysis demonstrating that four independently developed technology categories (AR wearables, nudification models, real-time video generation, and edge AI inference) could be combined into a system for ambient non-consensual intimate image synthesis (Gilly, 2026a). Eight weeks later, we documented that a single week of capability releases had compressed the feasibility timeline (Gilly, 2026b). In a companion paper, we analyzed how the same convergence stack enables therapeutic perceptual modification through regulatory pathways established by existing digital therapeutics (Gilly, 2026c). In a third companion, we showed that a single technology roundup video contained eight individually benign open-source tools that collectively constitute infrastructure for population-scale perceptual control (Gilly, 2026d).

These are not three separate problems. They are three instances of a single structural phenomenon: independently developed capabilities combining to produce emergent harm that is invisible to governance frameworks designed to evaluate capabilities in isolation.

The question we address in this paper is not "how do we govern perceptual modification" or "how do we prevent ambient nudification." Those are application-specific questions addressed in our companion papers. The question here is methodological: why did existing governance frameworks fail to detect these threats, and what kind of methodology could succeed?

1.2 The Component Evaluation Paradigm

Every major AI governance framework currently in operation shares a common structural assumption: the unit of evaluation is a single system, model, or application.

The EU AI Act assigns risk classifications to "AI systems" based on their intended purpose (European Parliament & Council, 2024). A facial recognition system is high-risk. A chatbot is limited-risk. A spam filter is minimal-risk. The classification applies to the system, evaluated in isolation.

The NIST AI RMF provides structured risk management for "AI systems" through four functions (Govern, Map, Measure, Manage; NIST, 2023). Each function operates on a specific system with defined boundaries, stakeholders, and deployment context.

The MIT AI Risk Repository catalogs risks associated with AI technologies, organized by domain and harm type (MIT FutureTech, 2024). Each entry describes a risk associated with a specific capability or application.

ISO/IEC 42001:2023 specifies requirements for establishing, implementing, and maintaining an AI management system (International Organization for Standardization [ISO], 2023). The management system is organized around specific AI products and services.

None of these frameworks provides a mechanism for asking: "What happens when Tool A, developed by Team A for Purpose A, is combined with Tool B, developed by Team B for Purpose B?"

This is the component evaluation paradigm: the assumption that evaluating individual components is sufficient to ensure system-level safety. We argue this assumption is false for AI capabilities with compatible interfaces, and that its falsity is not merely theoretical but empirically demonstrated by the case studies in our companion papers.

1.3 Why This Was Not Always a Problem

The component evaluation paradigm was adequate when AI capabilities were:

Siloed by modality (text models did not interact with image models did not interact with audio models). Each capability existed in its own ecosystem with distinct data formats, toolchains, and deployment environments.

Siloed by compute requirements. Running a state-of-the-art model required specialized hardware, significant engineering, and institutional resources. Integration was expensive.

Siloed by access. Models were proprietary, behind APIs, or available only through commercial platforms. Combining capabilities from different providers required negotiating access to each.

All three silos have collapsed. Modern AI development is multimodal by default; models routinely handle text, image, audio, and video. Compute requirements have dropped by orders of magnitude through distillation, quantization, and architectural innovation; the tools in our case study include

models as small as 4.8 megabytes and 140 megabytes running on consumer phones. Access is overwhelmingly open; HuggingFace alone hosts over 300,000 model repositories, and open-source release has become the default for academic research.

The result is that the combinatorial space of possible integrations has expanded from a tractable number of expensive, difficult connections to an effectively infinite number of cheap, easy connections. The governance paradigm designed for the former world is operating in the latter.

2. Formalizing Convergence Risk

2.1 Definitions

We define the following terms for precision.

A *capability* is a specific function an AI model or tool can perform: segment a person from video, generate a 3D environment from an image, clone a voice from ten seconds of audio, accelerate video generation by a given factor.

An *integration surface* is the interface at which two capabilities can be connected. Integration surfaces exist wherever output from one capability can serve as input to another. They may be explicit (documented APIs, shared file formats) or implicit (compatible data types, common intermediate representations).

Convergence risk is the potential for harm that exists at an integration surface but does not exist for either capability individually. The risk is emergent: it arises from the combination and cannot be detected by evaluating either component alone.

A *convergence threat* is a specific harmful outcome enabled by a specific combination of capabilities across a specific set of integration surfaces.

2.2 Properties of Convergence Risk

Convergence risk exhibits three properties that distinguish it from conventional AI risk.

The first is invisibility under component evaluation. By definition, convergence risk does not appear when capabilities are assessed individually. A person segmentation model has no harm profile that includes "enabling ambient nudification." A world generation model has no harm profile that includes "enabling perceptual reality replacement." These harm profiles exist only at the integration surface.

The second is combinatorial explosion. If there are N independently developed capabilities with compatible interfaces, the number of possible pairwise combinations is N(N-1)/2. For triples, it is N(N-1)(N-2)/6. The space of possible combinations grows polynomially for fixed combination sizes and exponentially if combination size is unbounded. With over 300,000 models on HuggingFace alone, exhaustive evaluation of all possible combinations is computationally infeasible.

The third is temporal distribution. The components of a convergence threat may be developed months or years apart, by teams that never communicate, in different countries, for entirely unrelated purposes. The threat does not "arrive" at a single point in time; it assembles gradually as independent development trajectories converge. There is no moment at which a single actor creates the danger.

2.3 Distinguishing Convergence Risk from Misuse Risk

Convergence risk is distinct from the misuse of a single system. Misuse involves applying a specific tool for an unintended harmful purpose: using a language model to generate phishing emails, using an image generator to produce NCII, using a voice cloner for fraud. Misuse can be addressed (imperfectly) by the tool's developer through safety training, content filtering, and use policies.

Convergence risk involves no individual tool being misused. Each tool is used for its intended purpose. The person segmentation model segments people. The world generator generates worlds. The voice cloner clones voices. The harm arises from composition, not from any single tool operating outside its design parameters.

This distinction matters for governance. Misuse-oriented interventions (content filtering, acceptable use policies, model safety training) address harms within a tool's scope. They cannot address harms that emerge between tools, because no individual tool's developer is responsible for or even aware of the convergence.

3. The Chemical Analogy

3.1 Reactive Chemistry as Precedent

Physical sciences faced a structurally analogous problem: individually safe chemicals can produce dangerous reactions when combined. The solution was not to evaluate chemicals more carefully in isolation; it was to develop a new discipline, reaction chemistry, that studies interactions between substances.

Material Safety Data Sheets (MSDS, now SDS) document the properties of individual substances, including incompatibilities: lists of other substances with which the chemical should not be stored or combined. These incompatibility lists are the product of systematic study of chemical interactions, not merely of individual chemical properties.

The Globally Harmonized System of Classification and Labelling of Chemicals (GHS) extends this by standardizing hazard communication across borders, ensuring that incompatibility information travels with the substance regardless of jurisdiction (United Nations, 2023).

Neither MSDS nor GHS eliminates chemical accidents. But they provide the methodological infrastructure for detecting reactive combinations before incidents occur. The existence of the discipline is itself the safety mechanism.

3.2 What AI Governance Lacks

AI governance has equivalents to MSDS (model cards, system cards) and GHS (the EU AI Act's classification system). What it lacks is the equivalent of reaction chemistry: a discipline dedicated to studying what AI capabilities produce when combined.

Model cards (Mitchell et al., 2019) document individual model properties: training data, intended use, performance benchmarks, known limitations. They do not document integration surfaces or combinatorial risks. A model card for MatAnyone2 would describe its segmentation capability, intended applications in video editing, and performance benchmarks. It would not (and within its

current scope, should not be expected to) document that its output is pipeline-compatible with nudification models or environmental generators.

System cards (Meta, 2023) extend model cards to deployed systems, documenting system-level properties. They are closer to what we need, but they still describe a single bounded system, not the space of possible combinations with external capabilities.

The gap is disciplinary, not informational. It is not that existing documentation is incomplete (though it is). It is that no discipline exists whose purpose is to study what independently developed AI capabilities produce when combined.

4. A Proposed Methodology: Integration Surface Analysis

4.1 Foundations

We propose Integration Surface Analysis (ISA) as a methodology for detecting convergence risk. ISA extends the Harm Blindness Framework (Gilly, 2025-2026) from product-level analysis to cross-capability analysis. Where HBF asks "who gets hurt" at decision checkpoints within a single development process, ISA asks "who gets hurt" at integration surfaces between independent development processes.

4.2 Step 1: Capability Mapping

The first step is to maintain a continuously updated map of AI capabilities, their input/output interfaces, and their compatibility with other capabilities.

This does not require mapping every model on HuggingFace. It requires mapping capability classes: person segmentation, environment generation, voice synthesis, video acceleration, and similar functional categories. Within each class, the relevant properties are input format, output format, latency, size/compute requirements, and deployment accessibility (open source, API, proprietary).

Existing catalogs like the MIT AI Risk Repository and PapersWithCode provide partial foundations for this mapping but are organized around individual capabilities rather than interface compatibility.

4.3 Step 2: Integration Surface Identification

For each pair of capability classes with compatible interfaces, identify the integration surface: what data flows from one to the other, what the combined output is, and what new functionality the combination enables that neither component provides alone.

Not all integration surfaces are concerning. Most are benign (text-to-speech combined with translation produces a multilingual speech system, which is useful and harmless). The challenge is filtering the vast space of possible surfaces to those warranting analysis.

We propose three filtering criteria.

The first is perceptual modification: does the combination enable modification of a human's sensory experience (visual, auditory, haptic) in a way that neither component enables alone?

The second is consent bypass: does the combination affect individuals who did not consent to or are unaware of the system's operation?

The third is scale amplification: does the combination operate at a scale (population, geographic, temporal) that exceeds either component's individual reach?

These criteria are not exhaustive. They are derived from the case studies in our companion papers and represent the convergence risks we have identified to date. A mature ISA methodology would develop a more comprehensive set of filtering criteria through systematic analysis of convergence incidents.

4.4 Step 3: Harm Analysis at Flagged Surfaces

For each integration surface that passes the filtering criteria, apply the HBF harm analysis:

Who is directly affected by the combined capability? Who is indirectly affected (through dependency, proximity, or social relationship with directly affected individuals)? What populations are affected who were not stakeholders for either individual capability? What consent mechanisms exist for newly affected populations? What is the minimum technical barrier to achieving the integration? What is the reversibility of the combined effect?

The output of this analysis is a convergence risk profile for the integration surface, comparable in function to an SDS incompatibility listing for chemicals.

4.5 Step 4: Monitoring and Alerting

Because the capability landscape changes continuously, ISA must operate as a monitoring function rather than a point-in-time assessment.

When a new capability is released (a new model on HuggingFace, a new paper on arXiv, a new tool on GitHub), it is mapped into the capability landscape and its integration surfaces with existing capabilities are evaluated against the filtering criteria. If new concerning surfaces are identified, harm analysis is triggered.

This monitoring function is precisely the type of task that can be partially automated using the same AI tools whose capabilities it monitors. Natural language processing can extract capability descriptions from papers and documentation. Compatibility analysis can be partially automated based on input/output specifications. The filtering criteria can be applied programmatically.

We note the irony and accept it: AI monitoring AI for convergence risk is itself a convergence of capabilities. The difference is that this convergence is intentional, governed, and directed toward harm prevention.

5. Limitations

5.1 Combinatorial Scale

The space of possible capability combinations is vast. Even with filtering criteria, the number of surfaces requiring analysis may exceed available analytical capacity. This is a scaling challenge, not a conceptual one, but it is real.

Prioritization heuristics can help: focusing on capabilities that are open source, small enough to run on consumer hardware, and recently released (since older capabilities are more likely to have been informally assessed by the research community). But any prioritization scheme will miss some combinations.

5.2 Dual-Use Ambiguity

Most capability combinations are genuinely dual-use: the same integration surface that enables harm also enables legitimate applications. Person segmentation combined with environment generation enables both perceptual manipulation and cinematic visual effects. Voice cloning combined with expressive synthesis enables both conversational impersonation and accessibility tools for people who have lost their voice.

ISA does not resolve dual-use ambiguity. It identifies the integration surface and the harm potential. Governance decisions about whether and how to intervene remain with policymakers, and those decisions require balancing harm prevention against legitimate use.

5.3 Implementation Gap

We have described a methodology. We have not built a system that implements it. The gap between methodology description and operational deployment is significant and requires engineering, funding, institutional support, and ongoing maintenance.

We note that the RSAIF Automator (Gilly, 2025-2026), a RAG-based system processing the HBF case database, provides a partial foundation for automated convergence scanning. Extending this system to monitor capability releases, map integration surfaces, and flag concerning combinations is an active area of development. But it is early-stage, and we do not overstate its current capability.

5.4 Adversarial Dynamics

If convergence risk profiles are published, malicious actors gain a roadmap for dangerous combinations. This is the same responsible disclosure tension we addressed in our original threat analysis (Gilly, 2026a, Section 1.6). We believe the benefits of systematic identification outweigh the risks of publication, because the components are already public and the combinations are discoverable by anyone who looks. Systematic identification gives defenders lead time.

6. Relationship to Existing Work

6.1 Compound AI Systems (BAIR/Banerjee et al.)

The Berkeley AI Research lab's framing of "compound AI systems" (Zaharia et al., 2024) and Banerjee et al.'s (2024) systematization of knowledge on compound AI threats are the closest existing work to our convergence risk framing. Both observe that modern AI results increasingly come from systems integrating multiple components, and both note that such systems can exhibit security properties not present in individual components.

Our contribution differs in scope and focus. Compound AI systems research examines intentionally composed pipelines within a single organization's product (an LLM plus a retrieval system plus a content filter). Our convergence risk framing examines unintentionally composable capabilities across

independent organizations with no shared governance. The compound systems literature asks "how do we secure this pipeline we built?" We ask "how do we detect this pipeline nobody built?"

6.2 Perception Engineering (LaValle et al.)

LaValle et al.'s (2024) formalization of perception engineering provides the control-theoretic foundation for understanding perceptual modification systems. Their framework models scenarios in which a producer alters the environment to alter the perceptual experience of a receiver, with formal treatment of adversarial and social engineering cases.

Our contribution is to connect this theoretical foundation to the concrete capability landscape, identifying specific tools and integration surfaces through which perception engineering becomes practically achievable, and to propose governance methodology for detecting such convergence.

6.3 Cognitive Infrastructure

A recent preprint on AI as cognitive infrastructure reframes AI systems as invisible infrastructures that condition what is knowable and actionable (Riva, 2025). Its emphasis on "adaptive invisibility" and the automation of relevance judgment aligns with our observation that perceptual modification systems are invisible to governance frameworks because they emerge between evaluated components.

6.4 Glasgow AR Ethics Network

Macpherson et al.'s (2023) AR Ethics, Perception, Metaphysics network explicitly treats AR as a new layer of infrastructure over public space, asking who has rights to place virtual content where. They note that AR can erase or re-render physical reality and may become as infrastructural as the internet. Their work provides the policy foundation; ours provides the methodological infrastructure for identifying which specific capability combinations make their concerns technically feasible.

7. Conclusion

We have argued that AI governance faces a structural gap: the inability to detect emergent harm arising from the combination of independently developed capabilities. This gap is not an oversight in any specific framework; it is a limitation of the component evaluation paradigm shared by all current frameworks.

We have proposed Integration Surface Analysis as a methodology for addressing this gap, extending the Harm Blindness Framework from product-level checkpoints to cross-capability integration surfaces. We have described a four-step process (capability mapping, integration surface identification, harm analysis at flagged surfaces, and continuous monitoring) and discussed its limitations.

The core claim of this paper is modest: the discipline we propose does not exist and should. The tools we need to build it are emerging. The case studies that justify it are documented in our companion papers. The cost of not building it is that we continue to inspect chemicals one at a time while the warehouse fills with reactive combinations.

We do not know how many dangerous convergences exist in the current capability landscape. We know that we found three, using informal methods, based on one week of releases covered in one YouTube video. A systematic methodology would, we believe, find more.

The question is whether AI governance develops the capacity to find them before they are exploited, or whether it waits for the explosion and then investigates the debris.

References

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

European Parliament & Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union.* EUR-Lex: 32024R1689.

Gilly, T. (2025-2026). *The Harm Blindness Framework: A practical application methodology for stakeholder harm prevention in technology development.* Real Safety AI Foundation. https://realsafetyai.org/framework

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

Gilly, T. (2026b). Compressed feasibility: A technical update to the ambient non-consensual synthesis threat model. *Forthcoming.*

Gilly, T. (2026c). We Happy Few: Therapeutic perceptual modification, regulatory pathway precedent, and the governance gap in cloud-rendered reality. *Forthcoming.*

Gilly, T. (2026d). The accidental stack: A case study in emergent perceptual control infrastructure from independently developed AI capabilities. *Forthcoming.*

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 Information technology: Artificial intelligence: Management system.*

LaValle, S. M., et al. (2024). From virtual reality to the emerging discipline of perception engineering. *Annual Review of Control, Robotics, and Autonomous Systems, 7,* 291-315. https://doi.org/10.1146/annurev-control-062323-102456

Meta. (n.d.). *System cards.* Meta Platforms, Inc. https://ai.meta.com/tools/system-cards

MIT FutureTech. (2024). *AI Risk Repository.* Massachusetts Institute of Technology. https://airisk.mit.edu

Mitchell, M., et al. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM. https://doi.org/10.1145/3287560.3287596

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Riva, G. (2025). Invisible architectures of thought: Toward a new science of AI as cognitive infrastructure. *arXiv preprint arXiv:2507.22893.*

United Nations. (2023). *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)* (ST/SG/AC.10/30, Rev. 10).

Macpherson, F., et al. (2023). *Policy and practice recommendations for augmented and mixed reality.* Centre for the Study of Perceptual Experience, University of Glasgow. Royal Society of Edinburgh Research Network Project. https://www.gla.ac.uk/research/az/cspe/projects/augmentedrealityethicsperceptionmetaphysics/

Zaharia, M., et al. (2024, February 18). The shift from models to compound AI systems. *Berkeley AI Research Blog.* https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

Word Count: Approximately 5,200 words (excluding references)

Target Venue: AI Safety / Science & Technology Policy journal; companion SSRN/SocArXiv preprint