The Accidental Stack: A Case Study in Emergent Perceptual Control Infrastructure from Independently Developed AI Capabilities

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

Abstract

We present a case study in emergent convergence risk, documenting how a single week of independent AI capability releases (March 2026) produced the component stack for a population-scale perceptual control system that no individual developer intended, designed, or recognized. We inventory eight open-source tools released or updated within a seven-day period: a person segmentation model (MatAnyone2), a real-time 3D world generator (InSpatio-WorldFM), a mobile 3D renderer (MobileGS), a video generation accelerator (Diagonal Distillation), a voice cloner (TaDa), a tagged speech synthesizer (Fish Audio S2), a visual effects transfer system (Effect Maker), and a multi-shot cinematic video generator (Shotverse). Each tool was reviewed in isolation by a popular technology channel and evaluated as a standalone capability with legitimate applications. None were identified as combinable into a unified system. We apply integration surface analysis, extending the Harm Blindness Framework from product-level checkpoints to capability-combination checkpoints, to demonstrate how individually benign tools become infrastructure for perceptual modification when composed. We argue this case study illustrates a structural gap in existing AI governance: no current methodology, including the EU AI Act risk classification, NIST AI RMF, or MIT AI Risk Repository, systematically scans for emergent harm at integration surfaces between independently developed capabilities. This gap is not an oversight in implementation; it is a limitation of the methodological paradigm.

1. Introduction

In March 2026, a technology YouTuber published a routine weekly roundup of new AI tools (AI Search, 2026). The video covered approximately fifteen new releases across image generation, video generation, 3D rendering, voice synthesis, robotics, and consumer applications. Each tool was presented individually, with a brief demonstration, a quality comparison to prior tools, and links to open-source code or hosted demos.

The presenter did not recognize, and had no reason to recognize, that a subset of the tools he reviewed constituted the component stack for continuous, adaptive, real-time modification of a human being's perceptual experience of physical reality.

This is not the presenter's failure. It is a structural feature of how AI capabilities are developed, published, and evaluated. Each tool solves a specific technical problem. Each is evaluated against benchmarks for that specific problem. No evaluation asks what the tool enables when combined with other independently developed tools solving adjacent problems.

We use this specific case, one week, one video, eight tools, to illustrate a general problem: existing AI governance methodologies evaluate capabilities in isolation and are therefore structurally incapable of detecting emergent convergence threats.

The analogy is chemical safety. Each chemical in a warehouse may be individually below its flash point. A safety inspector who evaluates each chemical separately and stamps "compliant" on the manifest will miss the fact that storing them adjacent to one another creates a reactive mixture. The warehouse passes every inspection and explodes anyway. The inspector was not negligent; the inspection methodology was inadequate to the risk class.

We argue that AI governance is currently running chemical-by-chemical inspections on a warehouse full of reactive capabilities.

2. The Stack

2.1 Inventory

We document the eight capabilities below, with their key properties for convergence analysis.

MatAnyone2 is a video matting model that separates people from backgrounds in real-time video with state-of-the-art quality, including under conditions of rapid motion and complex hair (Yang et al., 2026a). The model is approximately 140 megabytes. It is fully open source with a hosted demo. The capability it provides to the convergence stack is real-time isolation of individuals from their environments for downstream processing.

InSpatio-WorldFM is a real-time generative frame model that produces spatially consistent 3D environments from single images or text prompts, with persistent spatial memory (Zhang, X., et al., 2026). It runs at interactive frame rates on an RTX 4090. It is fully open source with a hosted demo. The capability it provides is generation of coherent alternative environments that can replace or overlay physical reality.

MobileGS is a Gaussian splatting renderer compressed to 4.8 megabytes that achieves 120+ FPS on a Snapdragon 8 Gen 3 phone (Du et al., 2026). It is fully open source. The capability it provides is real-time 3D scene rendering on mobile and wearable-class hardware.

Diagonal Distillation is a video generation acceleration technique achieving 270x speedup over baseline diffusion models, generating five-second video in 2.6 seconds with maintained quality over sequences up to five minutes (Liu et al., 2026). It is fully open source. The capability it provides is near-real-time video generation on consumer hardware.

TaDa is a text-to-speech and voice cloning model that produces natural-sounding speech from ten seconds of reference audio, with zero hallucination rate and the fastest synthesis speed among benchmarked competitors. Models are 4 to 9 gigabytes. It is fully open source with a hosted demo.

Fish Audio S2 is a text-to-speech model supporting inline control tags for pausing, emphasis, laughter, whispering, shouting, and other expressive qualities. It is fully open source.

Effect Maker (Tencent Hunyuan) transfers visual effects from one video to another, copying specific VFX and applying them to new subjects. Code and dataset are forthcoming (announced as open

source).

Shotverse generates multi-shot cinematic videos with smooth camera transitions and consistent character appearance across cuts. It was trained on real cinematic data. Code and models are forthcoming (announced as open source).

2.2 What Was Not in the Video

Two additional developments from the same period merit inclusion for completeness.

Google Maps introduced AI-powered immersive navigation featuring Gemini-generated 3D reconstructions of real environments, delivered to billions of phones (Daniel, 2026). This is not open source, but it demonstrates that AI-generated 3D environmental rendering is already shipping as a consumer product at global scale.

RL3D Edit enables text-prompt editing of 3D scenes: changing objects, characters, backgrounds, and styles within complete 3D environments (Wang et al., 2026). This provides the editing interface for modifying generated or captured 3D environments based on natural language instructions.

3. Integration Surface Analysis

3.1 Method

We extend the Harm Blindness Framework (Gilly, 2025-2026) from its standard application (evaluating a single product or system at four decision checkpoints) to integration surface analysis. An integration surface is the interface at which two independently developed capabilities can be connected, whether through API, file format compatibility, data pipeline, or shared intermediate representation.

At each integration surface, we ask the HBF's foundational question: "Who gets hurt?"

This analysis does not require that anyone has actually built the integrated system. The point is that the integration is technically feasible and that the harm potential is invisible when each capability is evaluated in isolation.

3.2 Integration Map

We identify the following primary integration surfaces and their harm implications.

Surface 1: MatAnyone2 + WorldFM (Person Extraction + Environment Replacement)

MatAnyone2 isolates a person from their environment. WorldFM generates an alternative environment. Combined, they enable the replacement of everything around a person with generated content while preserving the person's appearance and position. The person becomes the only real element in an otherwise synthetic scene.

Who gets hurt: The isolated individual, whose physical context is erased and replaced without their knowledge or consent. Bystanders who are removed from the scene entirely. Anyone whose relationship to the individual depends on accurate environmental context (caregivers, emergency responders, colleagues).

Individual evaluation would find: MatAnyone2 is a video editing tool with applications in filmmaking, content creation, and accessibility. WorldFM is a spatial intelligence research platform with applications in robotics, gaming, and XR. Neither evaluation would flag the combination.

Surface 2: WorldFM + MobileGS (Environment Generation + Mobile Rendering)

WorldFM generates 3D environments on desktop GPUs. MobileGS renders 3D content on phones. Combined, they create a pipeline where environments are generated in the cloud or on a tethered device and rendered locally on mobile or wearable hardware at interactive frame rates.

Who gets hurt: Users of AR wearables receiving cloud-generated environments, who cannot verify what is being rendered (the verifiability problem identified in Gilly, 2026b). Populations receiving culturally encoded "default" environments generated from biased training data.

Individual evaluation would find: WorldFM advances spatial intelligence research. MobileGS advances efficient rendering. Neither evaluation would flag the pipeline.

Surface 3: TaDa + Fish Audio S2 (Voice Cloning + Expressive Control)

TaDa clones any voice from ten seconds of audio. Fish Audio S2 adds expressive control tags (emphasis, laughter, whispering). Combined, they enable the generation of emotionally modulated speech in anyone's cloned voice, controllable at the word level.

Who gets hurt: Anyone whose voice is cloned without consent, for purposes ranging from social engineering to emotional manipulation. Individuals in the perceptual environment of a user wearing AR glasses, whose real speech could be replaced or augmented with synthesized speech in real time.

Individual evaluation would find: TaDa is a text-to-speech model with applications in accessibility, content creation, and localization. Fish Audio S2 adds expressive control for more natural synthesis. Neither evaluation would flag the combination as enabling real-time conversational impersonation within a perceptual modification system.

Surface 4: DiagDistill + Any Visual Pipeline (Acceleration + Everything)

DiagDistill is a horizontal accelerator that makes any diffusion-based video generation 270 times faster. It does not generate content; it makes generation fast enough to be real-time. It is the capability that converts laboratory demos into deployed systems.

Who gets hurt: DiagDistill does not directly harm anyone. It amplifies the harm potential of every other visual generation tool by removing the latency barrier between "this is possible in principle" and "this runs continuously on consumer hardware." It is the catalyst that makes the rest of the stack operational.

Individual evaluation would find: DiagDistill is an inference optimization technique with applications in efficient AI deployment. This evaluation is correct. It is also incomplete, because the optimization enables real-time operation of systems whose harm profile was bounded by their latency.

Surface 5: Full Stack (All Components)

When all components are connected: MatAnyone2 segments the user and their environment. WorldFM generates a replacement environment. MobileGS renders it on wearable hardware. DiagDistill makes generation fast enough for continuous operation. TaDa and Fish Audio S2 generate emotionally appropriate audio in cloned or synthetic voices. Effect Maker transfers visual styles and effects.

Shotverse provides cinematic consistency across changing viewpoints.

The result is a system capable of continuously rendering an alternative reality for the user: visually coherent, spatially persistent, auditorily complete, emotionally adaptive, and cinematically smooth. Every person in the user's environment can be visually modified or replaced. Every surface, object, and space can be regenerated. Every voice can be substituted. And the entire experience adapts in real time to maintain continuity.

No individual developer built this system. No individual tool evaluation would detect it. The system is emergent: it exists in the spaces between tools, at the integration surfaces, and nowhere else.

4. Why Current Governance Misses This

4.1 The EU AI Act

The EU AI Act (European Parliament & Council, 2024) classifies AI systems into risk tiers: unacceptable, high, limited, and minimal. Classification is based on the system's intended purpose and application context.

Every tool in our inventory would likely be classified as minimal or limited risk individually. MatAnyone2 is a video processing tool. WorldFM is a research demonstration. MobileGS is a rendering optimization. TaDa is a text-to-speech model. None meet the criteria for high-risk classification under Annex III, which focuses on biometrics, critical infrastructure, education, employment, law enforcement, and similar domains.

The Act does not provide a mechanism for evaluating what happens when minimal-risk tools are composed into a system that would, if built as a single product, warrant a different classification entirely.

4.2 NIST AI Risk Management Framework

The NIST AI RMF (National Institute of Standards and Technology [NIST], 2023) provides a structured approach to AI risk management organized around four functions: Govern, Map, Measure, and Manage. The framework is applied to specific AI systems and their deployment contexts.

The framework does not address the scenario in which no single AI system exists to evaluate. The perceptual control stack we describe is not a system; it is a collection of open-source tools that could be assembled into a system. No organization governs it. No deployment context can be mapped, because it has not been deployed. No measurements can be taken, because there is nothing yet to measure.

The NIST framework is designed for evaluating systems that someone has built and intends to deploy. It is not designed for scanning the space of possible combinations among existing tools. This is not a criticism of the framework's quality; it is a statement about its scope.

4.3 MIT AI Risk Repository

The MIT AI Risk Repository (MIT FutureTech, 2024) catalogs AI risks in a structured taxonomy. Risks are classified by domain, capability, and harm type.

The repository catalogs risks that have been identified by researchers. It does not systematically generate new risk hypotheses by examining capability combinations. A risk must first be recognized by a human researcher before it can be cataloged. The convergence stack we describe was not in the repository as of our review, because no researcher had identified it.

This is the fundamental limitation: cataloging known risks does not discover unknown ones. Convergence risks are, by definition, unknown until someone examines the integration surfaces. The repository is a map of explored territory, not a method for exploring new territory.

4.4 The Common Structure

The EU AI Act, NIST AI RMF, and MIT AI Risk Repository share a common structural feature: they evaluate AI capabilities (or systems, or risks) as discrete, bounded entities. This is appropriate for the risk classes they were designed to address. It is inadequate for convergence risk, where the threat is not in any entity but in the relationship between entities.

The analogy to chemistry remains apt. The periodic table catalogs elements. Material Safety Data Sheets describe individual chemicals. Neither tells you what happens when you combine them. That requires a different discipline entirely: reaction chemistry, which studies the interactions between substances rather than the substances themselves.

AI governance needs its reaction chemistry.

5. Toward Convergence Scanning

5.1 The Problem Statement

The governance gap we have identified can be stated precisely: there exists no systematic methodology for scanning independently developed AI capabilities for emergent harm at integration surfaces.

"Systematic" means it does not depend on an individual researcher happening to notice a dangerous combination. "Integration surfaces" means it examines the connections between capabilities, not the capabilities themselves. "Emergent" means the harm is not present in any individual component and cannot be detected by evaluating components in isolation.

We do not claim to have solved this problem. We offer preliminary observations on what a solution might require.

5.2 Requirements for Convergence Scanning

A convergence scanning methodology would need to satisfy several requirements.

First, it must operate on capabilities, not products. The threat exists before anyone builds an integrated system. Scanning must examine what tools can do and how those capabilities compose, not what someone has already assembled.

Second, it must be continuous. New capabilities are released daily. A point-in-time assessment becomes stale immediately. The scanning process must operate as a monitoring function, not a periodic audit.

Third, it must scale across domains. The convergence stack we documented spans computer vision, generative modeling, 3D rendering, speech synthesis, and consumer hardware. A scanning methodology limited to a single domain will miss cross-domain convergence by definition.

Fourth, it must tolerate ambiguity. Most capability combinations are benign. Some are concerning. A very small number are dangerous. The methodology must filter a vast space of possible combinations down to those warranting detailed analysis, without producing so many false positives that the system becomes unusable or so many false negatives that it misses real threats.

Fifth, it must be grounded in harm, not in capability alone. Two capabilities combining to produce a more capable system is not inherently concerning. What matters is whether the combined capability creates new pathways to harm that do not exist for either component alone. This is why we ground our integration surface analysis in the HBF question "who gets hurt" rather than in capability benchmarks.

5.3 Candidate Approach: HBF at Integration Surfaces

The Harm Blindness Framework (Gilly, 2025-2026) was designed for product-level analysis, asking "who gets hurt" at four decision checkpoints during development. We propose extending HBF from checkpoints within a single development process to integration surfaces between independent development processes.

For each pair of capabilities with compatible interfaces (shared data formats, compatible APIs, complementary pipeline stages), the extended framework asks:

What does the combination enable that neither component enables alone? Who is affected by the combined capability who was not affected by either component individually? What is the reversibility of the combined effect? What consent mechanisms exist for the newly affected population? And what is the minimum technical barrier to achieving the combination?

This last question is critical. If combining two tools requires a research team, significant engineering, and novel technical work, the convergence risk is lower (not zero, but bounded by the effort required). If combining two tools requires glue code that an LLM can generate in minutes, the convergence risk is immediate.

Every tool in our case study falls into the latter category. Open source, documented, with compatible data formats and standard interfaces. The glue code is trivial.

6. The Cyberpunk and We Happy Few Parallels

Science fiction has explored perceptual control through two complementary lenses that illuminate the governance challenge.

In *Cyberpunk 2077* (CD Projekt Red, 2020), braindance technology allows users to experience recorded sensory experiences of others, and various cybernetic modifications alter perception directly. The fictional model assumes perceptual control requires invasive hardware, cyberware implants installed by licensed technicians, corporate infrastructure, and black-market supply chains. Control is centralized because the technology is centralized.

In *We Happy Few* (Compulsion Games, 2018), the perceptual modification is pharmacological (Joy) and self-administered. Control is decentralized because anyone can take or refuse the drug. But social pressure makes refusal costly; "Downers" who refuse Joy are expelled from society.

The real convergence stack resembles neither model exactly. Like Cyberpunk, it involves technological rather than pharmacological modification. Like We Happy Few, it is distributed and self-administered. But unlike both, it is emergent. No corporation built braindance. No government distributed Joy. The stack assembled itself from components developed by independent teams pursuing independent objectives.

This is what makes it ungovernable under current frameworks. There is no Arasaka to regulate. There is no Joy factory to shut down. There are thousands of GitHub repositories, each one solving an interesting problem, each one passing every individual safety assessment, and the combination assembling itself in the spaces between them.

Fiction assumed perceptual control would be intentional, centralized, and recognizable. The reality is accidental, distributed, and invisible to every existing governance methodology. That invisibility is the finding of this paper.

7. Conclusion

We have documented a specific case in which eight independently developed, individually benign, open-source AI tools released in a single week constitute the component stack for a perceptual control system that no developer intended. We have applied integration surface analysis to identify harm potential at each connection point. We have shown that current governance frameworks (European Parliament & Council, 2024; NIST, 2023; MIT FutureTech, 2024) are structurally incapable of detecting this class of risk because they evaluate capabilities in isolation.

We do not argue that someone will build this specific system tomorrow. We argue that the components are available, open source, documented, and trivially combinable, and that no systematic process exists to detect this fact. If the specific combination we identified is the only dangerous convergence in the current capability landscape, we have gotten extraordinarily lucky. We do not believe we have.

The foundational question is not whether this particular stack is dangerous. It is whether AI governance will develop the methodological capacity to scan for convergence threats systematically, or whether it will continue to evaluate chemicals one at a time while the warehouse fills.

References

AI Search. (2026, March). AI maps, realtime 3D worlds, multi-shot videos, new TTS, new anime model: AI NEWS [Video]. YouTube.

CD Projekt Red. (2020). *Cyberpunk 2077* [Video game]. CD Projekt.

Compulsion Games. (2018). *We Happy Few* [Video game]. Gearbox Publishing.

Daniel, M. (2026, March 12). How we're reimagining Maps with Gemini. *Google Blog.* https://blog.google/products-and-platforms/products/maps/ask-maps-immersive-navigation/

Du, X., Jiang, Y., Ye, T., Han, S., & Liu, S. (2026). Mobile-GS: Real-time Gaussian splatting for mobile devices. In *Proceedings of the International Conference on Learning Representations (ICLR).* arXiv:2603.11531. https://xiaobiaodu.github.io/mobile-gs-project/

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

Liu, J., Liu, X., Mei, K., Wen, Y., Yang, M.-H., & Liu, W. (2026). Streaming autoregressive video generation via diagonal distillation. In *Proceedings of the International Conference on Learning Representations (ICLR).* arXiv:2603.09488. https://github.com/Sphere-AI-Lab/diagdistill

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

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

Wang, J., et al. (2026). Geometry-guided reinforcement learning for multi-view consistent 3D scene editing. *arXiv preprint arXiv:2603.03143.* https://github.com/AMAP-ML/RL3DEdit

Yang, P., Zhou, S., Hao, K., & Tao, Q. (2026a). MatAnyone 2: Scaling video matting via a learned quality evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* arXiv:2512.11782. https://github.com/pq-yang/MatAnyone2

Zhang, X., et al. (2026). InSpatio-WorldFM: An open-source real-time generative frame model for spatial intelligence. *arXiv preprint arXiv:2603.11911.* https://inspatio.github.io/worldfm/

Word Count: Approximately 5,000 words (excluding references)

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