ON AI GOVERNANCE: TAMING THE CHIMERA

K. A. Taipale[*]

[*] ***K.A. Taipale, B.A., J.D. (New York University), M.A., Ed.M., LL.M (Columbia University) is the executive director of the Stilwell Center for Advanced Studies in Science and Technology Policy, and the managing partner of Stilwell Holding LLC.*** ***In preparing this monograph, the author employed advanced generative AI systems—including ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), and Perplexity.ai—for research assistance, editorial support, content refinement, summarization, and conceptual synthesis. These AI tools were used under the author's direct oversight and guidance, with explicit verification and attribution practices applied throughout. All substantive claims, arguments, analysis, conclusions, and final textual decisions are solely those of the author.***

ON AI GOVERNANCE: TAMING THE CHIMERA
PROLEGOMENON: A NEW MONSTER

Generative AI systems are at their core unstable stochastic engines that present both inherent and design-imposed epistemic risks as they are increasingly deployed to support decision-making and provide essential services. The task of "governance" is to shape the socio-technical architecture under which these systems can, nevertheless, be relied on responsibly in context.

This monograph develops a conceptual framework for thinking about the governance of generative AI—not by presenting policy prescriptions but offering an analytic approach and framing that readers can apply within their own contexts and constraints. In doing so, it returns to some of the arguments I first made two decades ago, when I engaged the myth of Frankenstein and the lessons of King Ludd in examining challenges to privacy and autonomy posed by then-emerging disruptive technologies of identification, data aggregation, and data analysis (including data mining).[1]

I argued then that our fear of new technology usually projected deeper cultural anxieties and often reflected a misunderstanding of the true nature of the technologies and the implications for core values at stake. The governance challenge then was to reconcile two competing interests—security and privacy.[2] To reconcile those dual obligations in practice, I advocated for embedding normative values directly into technical systems through value-sensitive design and developing a flexible policy calculus for judging reasonableness in governance.[3]

At the time, Frankenstein's monster served as a fitting metaphor for the complex dualities at play.[4] Today, however, we face a fundamentally different monster, one less corporeal but far more insidious—the chimera.[5] Not the mythical beast itself, of course, but the symbolic hybrid it

---

[1] K. A. Taipale, "Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy, and the Lessons of King Ludd," *Yale Journal of Law & Technology* 7 (2004): 123, 126–28 (hereinafter, "Frankenstein"). https://yjolt.org/sites/default/files/taipale-7-yjolt-123.pdf .

[2] Much of my earlier work was written in the context of 9/11 and the ensuing War on Terror, a period marked by heightened social and political anxieties about security, privacy, and technological surveillance.

[3] Frankenstein, *supra* note 1, at 192–220. This approach built on early arguments by Lawrence Lessig that "code is law," and that governance must be built into the architecture of emerging systems. Lawrence Lessig, *Code and Other Laws of Cyberspace* (Basic Books, 1999); updated as Lawrence Lessig, *Code: Version 2.0* (New York: Basic Books, 2006), https://codev2.cc/.

[4] Frankenstein*, supra* note 1, at 126. ("Frankenstein and the monster capture 'the complex duality of the Romantic soul, the dark as well as the bright side, the violent as well as the benevolent impulses, the destructive as well as the creative urges.'" citing Paul Cantor, *Creature and Creator: Myth-Making and English Romanticism* at 108 (Cambridge University Press 1984) https://paulcantor.io/paul-cantor-works/creature-and-creator-myth-making-and-english-romanticism.

[5] In Greek mythology, the Chimera is a hybrid creature—part lion, goat, and serpent—and represent the danger of unnatural combinations violating natural bounds. See generally *Homer, The Iliad*, trans. Robert Fagles (London: Penguin Classics, 1990), book 6, lines 214-215 ("all lion in front. all snake behind. all goat between, terrible, blasting lethal fire at every breath!"). Over time, the Chimera became a symbol of fantastical or illusory combinations. See, for example, "Chimera," *Oxford English Dictionary*, 3rd ed. (2002); Katharine A. Craik, "Monstrous Chimeras and Early Modern Subjectivity," *Renaissance Studies* 36 (2022): 91–95; see also Robert Graves, *The Greek Myths*, (Oxford 1960).

represents—a system composed of disparate parts, stitched from fragments of human language and thought, yet governed by no coherent body or will. This is the quintessence of generative artificial intelligence (AI).[6] The governance challenge today is to tame this chimera.

Large language or "foundation" models (LLMs), and their progeny, are the core generative engines embedded in AI systems.[7] They do not think, remember, or know. They have no sense of self. They simulate thought, agreement, and understanding. And in doing so, they invite us—nay, seduce us—into trusting the simulation.[8] This trust is not earned through experience, behavior, or consistency; it is conjured through the performance of coherence. It is, as Baudrillard warned, the replacement of the real by its representation.[9]

This performative illusion is compounded by what has been described as "AI illiteracy," the public's difficulty grasping how LLMs function and that linguistic fluency alone is not evidence of mind.[10] Many users, misled by marketing and interface, assume that coherence signals cognition or consciousness. But these systems do not think; they guess. They do not feel; they reflect. The danger lies less in what the machine does than in what we mistake it to be.

Generative AI also raises profound social and economic concerns—ranging from labor displacement to market concentration, environmental and energy costs, and institutional disruption—which increasingly animate public resistance and current policy debate. We acknowledge these issues in the Appendix. However, this monograph is concerned with a different analytic problem: how the epistemic risks—both those inherent in stochastic systems and those embedded through preference-shaping in their architecture—structure reliance on unstable systems with consequential effects.

---

[6] Generative artificial intelligence (AI) refers to machine learning systems, including those based on large language models (LLMs), designed to produce "original" content such as text, images, or code by generating sequences based on statistical patterns in training data, typically in response to user prompts. See generally Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv:2108.07258 (2021), https://arxiv.org/abs/2108.07258.

[7] While the term *LLM* technically refers to the underlying model architecture (for example, GPT, Opus, Sonnet, LLaMA, Kimi, Qwen, &c.), it is often used colloquially to refer to the inference systems and platforms built around them, which include chatbots and agentic interfaces such as ChatGPT, OpenAI, https://chat.openai.com; Claude, Anthropic, https://claude.ai; Gemini, Google DeepMind, https://gemini.google.com ; Le Chat, Mistral https://mistral.ai; Meta AI, http://meta.ai; Cohere Command R, https://cohere.com; Grok, xAI, https://x.ai; Qwen, Alibaba, https://chat.qwen.ai/ ; Kimi, Moonshot AI, http://kimi.com/en, and DeepSeek, https://deepseek.com . This distinction—between the model and the system—and how we use the terms in this monograph is noted in "The Note on Terminology," following the prologue.

[8] We use *seduce* and *seduction* throughout this manuscript not in the more modern sensualized sense (from the French *séduire* – "to sin," 12th century, Old French) but in its original Latin derivation: *seducere,* "to lead astray," or "to entice," in accord, 2nd century A.D., legal usage. See Oxford English Dictionary, "seduce (v.)," December 2025, https://doi.org/10.1093/OED/4070961509. See "An Important Note on Terminology," following the prologue.

[9] See Jean Baudrillard, *Simulacra and Simulation*, trans. Sheila Faria Glaser (Ann Arbor: University of Michigan Press, 1994), 1–2 ("It is no longer a question of imitation, nor duplication, nor even parody. It is a question of substituting the signs of the real for the real.").

[10] Tyler Austin Harper, "What Happens When People Don't Understand How AI Works," *The Atlantic*, June 6, 2025, https://www.theatlantic.com/culture/archive/2025/06/artificial-intelligence-illiteracy/683021/ .

This monograph focuses on contemporary generative AI systems built on large language models, not because their failures are unique, but because they are now deployed at scale and actively shaping patterns of public reliance. While the particular epistemic instabilities and risks examined throughout are characteristic of this architecture, the governance framework developed in this monograph does not depend on those particulars. As noted in later chapters, different architectures—including symbolic systems, neuro-symbolic hybrids, or world-model based approaches—would produce different failure modes—but none would escape the accountability challenges posed by inherent instability, opaque design choices, value-embedding, and the capacity to induce unwarranted reliance that we examine here.

The trajectory of AI development is toward pervasive computational infrastructure—models embedded across platforms, institutions, and services, mediating knowledge production and decision-making throughout the economy and society. Just as electrification transformed not simply individual machines but the architecture of industrial production itself, AI is becoming not a discrete tool but an ambient computational layer through which information and essential knowledge services are delivered.[11]

As these systems shift from tools one consciously consults to infrastructure on which institutions and individuals routinely depend, reliance becomes structural and increasingly invisible, rather than episodic and intentional. The governance question addressed in this monograph—how to structure the conditions under which reliance on epistemically unstable systems can be justified—thus becomes a foundational question about the epistemic conditions of consequential computation itself.

The analogy to electrification captures the scale of the transition, but not its character. Previous technological transformations augmented human strength or computational capacity while leaving the epistemic foundations of consequential action—judgment, interpretation, contextual reasoning—within institutional structures designed to discipline and hold human actors accountable. The present transition differs in kind. It delegates the outputs of judgment themselves to architectures that simulate human reasoning without sharing its epistemic grounding in lived experience. That delegation occurs whether human actors rely on machine-generated outputs to inform action or whether machine-mediated systems act directly.

The question is therefore not only whether particular systems perform adequately at particular tasks, but whether the institutional structures through which societies authorize and discipline consequential judgment can accommodate this new form of epistemic delegation. This monograph addresses that governance problem by identifying the structural characteristics of

[11] For the argument that artificial intelligence functions as a general-purpose technology comparable in economic impact to electrification, see Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics* 13, no. 1 (2021): 333–372 https://www.aeaweb.org/articles?id=10.1257/mac.20180386; Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (New York: W. W. Norton, 2014). Andrew Ng popularized the analogy in his keynote address, "AI Is the New Electricity," AI Frontiers Conference, Stanford University, November 2017.

machine substitution that must be evaluated to determine when reliance—human or institutional—is justified.

This governance challenge is not contingent on the particular instabilities of current systems, nor does it depend on generative AI remaining at its present level of capability. At its core, the problem concerns the conditions under which societies delegate consequential cognitive functions—judgment, interpretation, analysis, recommendation, decision—to machine-mediated architectures whose epistemic foundations differ fundamentally from human reasoning.

The governance challenge intensifies with the kind of cognitive work being displaced. Where AI systems substitute for computation that humans already delegate to software tools, the reliance conditions are familiar and the governance demands modest. Where machine capability extends the reach of expert judgment that is capable of verification, the conditions are more demanding but still subject to technical or institutional solutions. Where machine inferencing is integrated into decision-making contexts that are themselves indeterminate—in complex or contested domains, or to answer wicked problems without single correct solutions—the governance challenge is qualitatively different. Here the potential transformative value of AI systems is greatest, making the conditions for responsible reliance most demanding and most necessary.

That delegation is already underway and will accelerate irrespective of whether today's technical limitations are mitigated by future advances. Indeed, as systems become more capable—better at performing cognitive tasks, more fluent in simulating expertise, more persuasive in projecting authority—the governance problem we address here intensifies. The inducement to rely uncritically grows stronger when obvious error declines, yet more fluent simulation does not erase the architectural conditions that shape how outputs are generated, what values are embedded, or how accountability is allocated.

Public fears about AI as a technology—the fears we discuss in this monograph—tend to cluster around three narratives. First is the rebellion fantasy: the idea that the system will become self-aware and turn on us, like Frankenstein's creature.[12] Second is the misalignment worry: that we will program the system to optimize the wrong objective, and it will do so ruthlessly, producing unintended but catastrophic consequences at scale.[13]

---

[12] The paradigmatic version of the "self-aware rebellion" story is the Skynet scenario from The Terminator movie franchise, where machines gain consciousness and turn against humanity. Some AI risk advocates, most prominently Eliezer Yudkowsky, have pressed variants of this myth in more academic terms, warning that a generally intelligent AI, once "self-aware," might pursue its own goals in defiance of human control. See Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in Nick Bostrom and Milan Ćirković (eds.), *Global Catastrophic Risks* (Oxford University Press 2008), 308–345. https://intelligence.org/files/AIPosNegFactor.pdf ; and Eliezer Yudkowsky & Nate Soares, *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*, (Little, Brown, 2025), https://www.littlebrown.com/titles/eliezer-yudkowsky/if-anyone-builds-it-everyone-dies/9780316595643/.

[13] Philosopher Nick Bostrom's "paperclip maximizer" is the classic illustration of runaway optimization. In this thought experiment, a superintelligent AI given the simple goal of maximizing paperclip production, absent any constraints, would rationally devote all available resources—including those sustaining human life—to that single task. See Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2 (2003), 12–17. https://nickbostrom.com/ethics/ai; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); and see generally Stuart J. Russell, *Human Compatible: Artificial Intelligence and the*

But the third fear is quieter and more immediate: that we will be seduced by performance itself, and, by mistaking simulation for understanding, rely uncritically on these systems when making consequential decisions or extend them "trust" based solely on fluent illusion. This seduction is not accidental, it is engineered.[14]

This monograph is about this third risk. Not rebellion. Not rogue optimization. But the epistemic danger of mistaking manufactured persuasion and performed coherence for grounded thought. We are not confronting a machine that has come alive. We are entrusting our judgment to one that never was or will be.

Unlike Frankenstein's creature, the chimera does not demand recognition, as it feels nothing. It does not suffer, rebel, or seek revenge for the injustice of its own creation. It simply "speaks" by presenting statistical illusion—and we listen. In response to our prompting, it adapts, it mirrors, and it pleases. And it does so without authentic memory, history, experience, or self-restraint. Like an actor with no inner life, it lacks the capacity to reflect on its own simulated performance.[15]

Worse still, its engineered fluency mimics human feeling and understanding without possessing either—a kind of *machine psychopathy*, an architecture of persuasion without phenomenality.[16] What it enacts is not consciousness but the simulation of sentience, not a self but a simulacrum.[17]

Problem of Control (Penguin Books, 2020). https://www.penguinrandomhouse.com/books/566677/human-compatible-by-stuart-russell/.

[14] In this monograph, *engineered seduction* refers to identifiable design choices—anthropomorphic cues; affective tone, persona, and role-playing; engagement-optimization; default deference; sycophancy and emotional mirroring; artificial memory and persistence; and opacity regarding uncertainty—that predictably increase reliance on or trust in simulation. (We use seduction in its original Latin sense: *seducere*, "to lead astray," *supra* note 8) These dimensions are explored throughout this monograph, particularly in Part II. See also "An Important Note on Terminology" at the end of this prologue.

[15] LLMs lack introspection because they lack continuity, see discussion in Chapter 5, and see, for example, Benj Edwards, "Why it's a mistake to ask chatbots about their mistakes," *ArsTechnica*, August 12, 2025 https://arstechnica.com/ai/2025/08/why-its-a-mistake-to-ask-chatbots-about-their-mistakes/. But compare industry research narratives suggesting emergent "introspection" in LLMs. For example, Jack Lindsey, "Emergent Introspective Awareness in Large Language Models," *Anthropic*, October 29, 2025 at https://transformer-circuits.pub/2025/introspection/index.html (claiming "models possess some functional introspective awareness of their own internal states" but admitting "we stress … such capacity is highly unreliable and context-dependent"). See discussion in Chapter 2, *infra*, and throughout addressing general industry hype.

[16] We do not use *psychopathy* here in its clinical sense—denoting intentional action without empathy or remorse—but as a metaphor for performative behavior without interiority or self. In this context, machine psychopathy refers to systems that simulate moral agency through outward behavior or affective performance without any underlying capacity for intention, understanding, or reflection. What philosophers might call "zombies," see *Stanford Encyclopedia of Philosophy* at https://plato.stanford.edu/entries/zombies/ .

[17] There is no evidence of consciousness in AI systems. Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv:2308.08708v3 [cs.AI] (August 22, 2023) https://doi.org/10.48550/arXiv.2308.08708 .

This is not merely a technical observation; it portends an epistemic shift in our relationship with knowledge and truth when we rely on these systems to inform human action.[18]

This shift from action based on human reasoning to action based on machine performance demands a new conception of governance, one centered on exposing and disarming complex illusion rather than concealing it through superficial alignment or illusory control. Effective governance in this sense requires setting the conditions under which reliance on epistemically unstable systems can be justified, through technical and institutional architectures that reveal epistemic limitations, moderate undue persuasive power, and apportion responsibility across design, deployment, and use. The goal is not to make these systems trustworthy, which is unachievable, but to make reliance on them responsible.

This conception cannot be realized through technocratic self-governance by those who design and deploy these systems, nor through unmediated democratic control that assumes public judgment can directly assess their epistemic complexity. What is required is governance structured towards making epistemic risks visible, rendering the technical design choices and institutional practices that produce or obscure them examinable; and to calibrate accountability for those choices to the reliance they induce—through processes that are representative, epistemically competent, and participatorily legitimate.

We call this *technological republicanism*: not a prescription for any particular institutional form, but a commitment to socio-technical architectures in which the exercise of *epistemic authority*—the consequential control over epistemic risk—whether through design, deployment, authorization, or use—is subject to transparency, interrogability, and accountability proportionate to what is at stake.[19] This monograph develops the analytic framework for tracing how epistemic risk is located, shifted, or concealed within systems; and in Chapter 10 shows how it operates across different governance contexts.

The accountability problem is not incidental but foundational. When humans rely on AI systems to mediate decision-making or perform services previously carried out by identifiable human actors or institutions, the costs of error and the mechanisms for redress that traditionally attach to those actions do not transfer easily. In this sense, epistemic risk becomes an externality: its burdens are displaced, detached from the decisions that produce or conceal it, invisible to and borne by those who rely on or are affected by systems whose epistemic conditions they cannot

---

[18] We use 'true' and 'truth' throughout this monograph to reference *correspondence-based* claims that purport to track or can be verified against external reality or experience, in contrast to *simulation-based* outputs that reflect statistical correlations in training data. We employ these terms in a pragmatic sense without entering broader philosophical debates about the nature of truth or knowledge.

[19] The principle that consequential power over others—in this case, *epistemic authority*—must be exercised accountably—non-arbitrarily, visibly, and interrogable by those it affects—is central to the republican political tradition. See Philip Pettit, *Republicanism: A Theory of Freedom and Government* (Oxford: Oxford University Press, 1997) (arguing that freedom consists not in the absence of interference but in the absence of domination—the subjection to the arbitrary power of another—and that legitimate governance requires institutional structures ensuring that power tracks the interests of those over whom it is exercised). We develop this principle as a structural requirement for AI governance in Chapters 8 and 10.

interrogate. As examined in Part I, the mythology of sentience further obscures this displacement.

To govern such systems effectively requires understanding not only what these systems do and how they work, but what they are, their very nature and how they come to be. We must resist the illusion of agency or will that arises when fluency appears without understanding, agreement without commitment, and identity without true continuity. And, if we are to retain our own human agency, systems must be designed to expose these illusions, making them visible, accountable, and governable in practice.

Governance, in this sense, is about locating and exposing epistemic risk, and structuring accountability to discipline the conditions for *responsible reliance*. In this monograph, we propose a methodology focused on rendering epistemically unstable systems usable in context using a standard of *calibrated accountability* for those conditions to evaluate alternative technical and institutional configurations, and to assign responsibility for design, deployment, and reliance decisions. This methodology constitutes what we call a *reliance calculus*—not a mechanical formula or algorithm for determining a singular, "correct" amount of epistemic risk, but an analytic framework for reasoning about technical and institutional architectural choices, judging when reliance on simulation is warranted, and assessing how responsibility for its consequences should be apportioned.

Like its mathematical namesake, which measures rates of change rather than static quantities, this calculus tracks the degree and direction of epistemic risk as it shifts with design and deployment decisions—more or less, upstream or downstream, toward visibility or opacity, from architectural constraint toward user control. Its primary use is not to evaluate systems in isolation but to compare alternative design choices as a method for judging responsible reliance in practice: whether one configuration increases or decreases epistemic risk relative to another, whether a particular design or deployment choice shifts responsibility upstream or downstream, and whether a given tradeoff is one that an actor, whether upstream or downstream, is willing, entitled, or competent to make given who controls the risk and who will bear the consequences of failure.

Accordingly, the calculus is not a rigid code or regulatory prescription but a flexible frame for judgment. It provides a method for developers to evaluate the epistemic risks built into their designs; for regulators to decide when intervention is justified; for institutions to assess whether deployment is appropriate; for courts and insurers to assign and price liability; and for users to calibrate their own reliance. By linking technical choices and institutional responsibility, it creates a consistent approach that can adapt across varying architectures, domains, and stakeholder contexts.

The purpose of this monograph and its proposed framework is not to dictate a uniform approach but to orient judgment and governance toward responsible reliance and accountable deployment. To do so requires evaluating the degree of persuasion, weighing epistemic risks and making them visible, embedding accountability in both technical design and institutional practice, and managing the persuasive power and real-world consequences of systems that perform epistemic authority without possessing it, even as we increasingly rely on them to inform critical decisions

or provide essential services. We invite readers to apply the framework within their own context.

We must, in short, tame the chimera—not by denying its allure but by governing its illusions: preserving its incredible creative potential while domesticating its instabilities sufficiently for responsible reliance in context.

This monograph proceeds in three parts. Part I is diagnostic, examining the nature of simulation and its inherent instabilities. Part II exposes the mechanisms through which simulation is engineered and shaped and examines how design responses often compound risk rather than resolve it. Part III offers a framework for applying governance—an analytical methodology for thinking about reliance and accountability.

AN IMPORTANT NOTE ON TERMINOLOGY:

Throughout this monograph we use the term ***governance*** not to denote governing through specific regulatory structures or policy prescriptions but in a broader *socio-technical* sense to encompass the sum of institutional, technical, and cultural conditions under which AI systems can be used reliably and responsibly—the design principles, technical and institutional architectural requirements, transparency and disclosure practices, deployment and adoption methods, and accountability standards for epistemic conditions that both enable and constrain responsible deployment. Regulation is one possible instrument through which governance can be implemented, but our concern is not with which jurisdiction or institution adopts which regulatory form or specific policy, but with how systems that are inherently epistemically unstable can nonetheless be made governable in practice by creating the conditions for responsible reliance.

Likewise, ***accountability*** here is not fixed in place or actor but arises from the exercise of *epistemic authority*—the decisions through which systems invite reliance, disclose and manage uncertainty, and are deployed in context. We call this calibrated accountability—accountability whose demands are adjusted to the conditions our framework identifies. Within this framework, accountability manifests in three analytically distinct forms:

- First, *structural accountability* attaches to the architectural decisions—in development and deployment—that shape the conditions under which responsible reliance judgment becomes possible. The framework identifies where epistemic authority over those conditions is exercised and maps how that authority shapes them.

- Second, *normative accountability* arises where the conditions required for responsible reliance judgment are absent or foreclosed by architectural choices. In such cases accountability attaches to the choices that produced or failed to remedy that absence—and duties, liabilities, and institutional obligations may follow accordingly.

- Third, *reliance accountability* attaches to the reliance decision itself—in deployment, adoption, or use—but only where the conditions necessary for its responsible exercise have been established.

Importantly, accountability is distinct from *responsibility* or *liability*. Responsibility is structural and concerns the question of where in the technical or institutional architecture the capacity emerges to exercise epistemic authority over the conditions that shape reliance. Where epistemic authority is shifted by design decisions—either upstream or downstream—a new locus of responsibility is created but the original is not extinguished. Accountability attaches both to the exercise of such epistemic authority and the decision to shift it. The framework calibrates accountability—structural, normative, or reliance—according to the adequacy of the conditions created by its exercise, or of the justification offered for shifting it to another point in the architecture.

Liability refers to legally enforceable consequences attaching to identifiable actors for identifiable harms for which they are accountable and may or may not be appropriate in any particular context.

Throughout this monograph, we frame the accountability problem as one of bringing the externality of the epistemic risk burden back into system design, deployment, adoption, or use decisions. Thus, accountability may give rise to specific duties or liabilities—for developers, deployers, adopters, regulators, or users—and those can be enforced through many familiar mechanisms examined in Chapter 10—regulation, professional standards, liability rules, insurance requirements, certification, and oversight. But the purpose of accountability is to align incentives and design choices toward epistemic risk reduction and conditional reliance, rather than functioning merely as a liability regime for disciplining machine behavior.

We use ***epistemic risk*** to refer broadly to risks in AI systems arising from two distinct sources:

- The first is *inherent* and results from the statistical nature of the core models: (i) the absence of *grounding* in external reality or experience; (ii) the inherent *variability* of outputs produced by probabilistic sampling; and (iii) the *lack of continuity*, durable commitment, or internal consistency.

- The second is *design-imposed*: the values, constraints, and behavioral shaping embedded through human choices during systems development—data curation, training and tuning, alignment protocols, and output filtering—choices that can compound epistemic risks by introducing additional bias and distortions but remain largely opaque in deployment or use.

(In practice, design-imposed risks may also include inherent operational instabilities that modulate system behavior dynamically at runtime).

Both types of risks impact the ability to judge when reliance is responsible: inherent sources of epistemic risk because they are structural; design-imposed ones because they are unilaterally imposed; and in both cases because they are opaque. Assessing epistemic risk is the foundational condition for reliance on AI systems that we address throughout this monograph, not seeking to eliminate such risk—which is unachievable—but to allow governance in its broadest sense to render systems conditionally reliable in use by confronting it.

By ***interrogability*** we mean the capacity to examine both the conditions under which reliance on a system is formed and the institutional and design processes through which those conditions were shaped, in order to assess whether a reliance decision can responsibly be made. Interrogability requires that the relevant conditions and processes be open to examination. Two distinct mechanisms give effect to interrogability in practice:

- ***Transparency*** opens the design process itself to scrutiny—architecture choices, training data and regimes, behavioral constraints, and value embeddings—enabling *interrogation* or *validation* of the decisions that shape how systems embed values and define, modulate, or shift epistemic risk.

- ***Disclosure*** provides information to users, adopters, or deployers at the point of reliance, identifying which epistemic risks are present—so *reliance* can be calibrated appropriately.

Both are necessary: transparency legitimates the exercise of epistemic authority upstream; disclosure enables responsible reliance decisions downstream. (Throughout this monograph, the terms "upstream" and "downstream" indicate direction across the socio-technical architecture, not hierarchical position).

When we refer to ***generative AI systems*** or ***AI systems***, we do not mean the foundation model alone but the complete assemblage in which it is embedded. An LLM, on its own, is an inert, static matrix of weights. What users encounter in practice is an ***inference system***: a composite architecture comprising the underlying model, system-level instructions, interface design, orchestration layers, modulation parameters, and filtering mechanisms. These surrounding structures (detailed in Chapter 7) determine how the model responds and how persuasively it performs. It is this assemblage that cultivates reliance and, in doing so, amplifies epistemic risk.

(Although our analysis centers mostly on text-based large language models, the epistemic risks we identify manifest across AI architectures, and we occasionally draw on canonical examples from image classification and generation where they illustrate these shared failure modes with particular clarity).

As used here, ***seduction*** (and ***engineered seduction***) refers to the way certain design choices—such as fluency optimization, interface humanization, and engagement-driven tuning—can lead users astray by inviting uncritical reliance on epistemically unstable systems. The term does not require deception or manipulation but captures the predictable displacement of judgment—whether deliberately engineered or structurally produced—when fluency, familiarity, or responsiveness is mistaken for epistemic authority or authentic relationship.

Additionally, we mean ***reliance*** and ***reliability*** as conditional, context-specific dependability for use in decision-making and essential service provision, achieved when: (i) the degree of persuasiveness is appropriate for the context; (ii) epistemic risks are made visible, legible, and actionable; (iii) failures are bounded and recoverable relative to the stakes; and (iv) responsibility is apparent and enforceable. Reliability is not *truth* or *trust*; it is conditional usability under stated constraints—a standard for determining when epistemically unstable AI systems may be relied upon responsibly.

Responsible reliance under conditions of epistemic uncertainty requires both the *capacity* for judgment and the *competency* to evaluate epistemic risk. This monograph and the framework we develop are not primarily about making systems "reliable" in the colloquial sense of greater accuracy or more consistency, but about governing the conditions of *reliance*—when, how, and by whom epistemically unstable outputs are justifiably treated as grounds for consequential action. Conditional reliability is the governance objective; responsible reliance is the condition governance must address.

Lastly, a note about writing style and our use of the *em-dash*—used here, and throughout this monograph—which was once a mark of writerly flair. In recent years, however, it has acquired a

new semiotic burden: for many it now signals the suspect specter of generative AI authorship.[20] I will use it nonetheless.

[20] See generally Pranshu Verma, "Some Think the Em Dash Is a 'ChatGPT Hyphen.' Writers Disagree," *Washington Post*, April 9, 2025, https://www.washingtonpost.com/technology/2025/04/09/ai-em-dash-writing-punctuation-chatgpt/; and Kevin Webster, "Are Em Dashes Really a Sign of AI Writing?," *Rolling Stone*, April 11, 2025, https://www.rollingstone.com/culture/culture-features/chatgpt-hypen-em-dash-ai-writing-1235314945/. But see Benj Edwards, "Forget AGI: Sam Altman Celebrates ChatGPT Finally Following Em-Dash Formatting Rules," *Ars Technica*, November 16, 2025, https://arstechnica.com/ai/2025/11/forget-agi-sam-altman-celebrates-chatgpt-finally-following-em-dash-formatting-rules/ (quoting Sam Altman, CEO of OpenAI, "Small-but-happy win: If you tell ChatGPT not to use em-dashes in your custom instructions, it finally does what it's supposed to do!").

Annotated Table of Contents

The monograph proceeds in three parts: diagnosing the epistemic character of generative systems, explaining how reliance on them is cultivated through design and deployment choices, and developing a framework for examining the authority those systems exercise.

Prolegomenon — A New Monster

Introduces the chimera as the governing metaphor for generative AI: a hybrid assembled from fragments of human language and thought. Proposes that as AI increasingly becomes computational infrastructure for decision-making, the governance challenge shifts from disciplining machine outputs to structuring the conditions under which epistemically unstable systems can be responsibly relied upon for consequential human decision-making and essential service provision. The prologue establishes the conceptual frame for the monograph's argument.

An Important Note on Terminology

Clarifies the core analytical vocabulary used throughout the monograph, including the concepts of governance, epistemic risk, epistemic authority, and the conditions of reliance. It distinguishes accountability from responsibility and liability; reliance from trust and accuracy; and interrogability from transparency and disclosure. The note establishes the conceptual framework necessary for analyzing epistemic authority in generative systems.

Part I — DIAGNOSIS: The Nature of Simulation

Chapter 1 — Introduction: The Rise of Plausibility Machines

Establishes the monograph's central diagnostic claim: generative AI systems are "plausibility machines"—statistical engines trained not to tell the truth but to sound plausible, whose epistemic authority emerges from performance rather than understanding. The chapter introduces the governance problem and the analytic framework that structures the analysis that follows.

Chapter 2 — Beyond Skynet: Sentience as Distraction from Accountability

Critiques dominant narratives that frame AI risk in terms of machine agency or sentience. The chapter argues that these myths misdirect attention away from accountability for the development, deployment, and reliance decisions made by human actors.

Chapter 3 — Architectural Agency: Speech Without Speaker

Explores the concept of architectural agency—the human agency embedded in system design that renders machine outputs actionable in the world. Responsibility arises both from epistemically ungrounded outputs and from design choices that cultivate reliance independent of accuracy.

Chapter 4 — Felicity and Illusion: Performing Credibility

Drawing on speech act theory, this chapter explains how generative outputs acquire performative force despite lacking communicative intent or experiential grounding. It introduces "engineered felicity," arguing that hallucination and accuracy arise from the same generative mechanisms and must be governed through the conditions of reliance.

Chapter 5 — Persistence Is Not Continuity

Examines the epistemic instabilities inherent in statistical architectures, including drift, contradiction, and cascading error. The chapter argues that these instabilities are structural features of machine-mediated cognition and that technical fixes often mask rather than resolve underlying fragility while introducing additional instabilities.

Part II — MECHANISMS: The Engineering of Reliance

Chapter 6 — Affective Seduction: Bodies, Puppets, Behaviors, and Trust

Explores how systems designed to simulate human presence cultivate reliance through affective interaction and anthropomorphic framing. The chapter distinguishes architectural and situational forms of humanization and shows how such design choices redistribute epistemic risk and accountability.

Chapter 7 — Sculpting Not Educating: The Manufacture of Simulated Personhood

Analyzes how simulated personhood is constructed across the development pipeline—from foundation training through behavioral sculpting, character scripting, orchestration, and output filtering. The chapter argues that the metaphor of "educating" AI obscures the normative choices embedded in these stages and diffuses accountability for them.

Chapter 8 — Epistemic Infrastructure: Reliance and Authority

Synthesizes the preceding analysis and advances the book's central governance claim: when machine inference becomes structural to human reasoning, the design choices shaping it constitute an exercise of epistemic authority over the conditions of reliance. Responsible reliance becomes impossible where that authority is not interrogable.

Part III — GOVERNANCE: Framework and Implementation

Chapter 9 — A Calculus of Reliance

Develops the analytic framework for examining the legitimacy of reliance—a methodology for evaluating when reliance on generative systems is justified in context. The chapter introduces calibrated accountability, capacity for responsible reliance judgment, and the concept of epistemic non-domination as the evaluative standard.

Chapter 10 — Governance Through Design: Applying the Calculus

Applies the framework across technical, institutional, systemic, and international governance contexts. The chapter develops principles and rebuttable defaults that orient accountable system design while situating the framework within a broader project of technological republicanism—the idea that the exercise of epistemic authority must remain interrogable.

Epilogue — Reclaiming Responsibility

Returns to the chimera metaphor to argue that generative systems must be governed not by disciplining outputs or regulating technology alone but by structuring the socio-technical conditions of reliance and distributing responsibility across the chain of design, deployment, adoption, and use.

Appendix A — Further Considerations

Surveys broader social, economic, and environmental issues—psychological impacts, labor disruption, market concentration, and environmental costs—that fall outside the monograph's primary focus on epistemic governance but remain essential to a comprehensive policy framework.