PART III

THE COST OF POLICY HALLUCINATIONS

Myths are not harmless fictions; they are the source code of flawed governance. When policies are built on fantasy, they become "policy hallucinations"—laws that govern shadows while real problems grow. This part takes stock of the real-world damage: the billions wasted on policy phantoms, the governance whiplash from hype and panic, and the democratic abdication as we cede control to unelected prophets and digital scapegoats.

It's time to examine the cost of our illusions.

CHAPTER 7

THE WRECKAGE OF MYTH-BASED GOVERNANCE

HOW IMAGINARY PROBLEMS PRODUCE REAL DAMAGE

The wreckage of myth-based governance manifests in three concrete forms: policy distractions that chase sci-fi fantasies, governance whiplash resulting from panic and hype, and the democratic abdication that leaves us unprepared.

Chasing Policy Phantoms, Wasting Billions

Legislating for Merlin

The first major cost is policy distraction: myths cause us to focus on the wrong problems. The Magic myth, by portraying AI as a human-like autonomous entity possessing agency comparable to that of human moral actors, raises complex legal and philosophical questions that actively misframe the challenges we face.[1] Policymakers become fixated on hypothetical questions about AI rights, consciousness, and moral status while overlooking the immediate, tangible harms created by

current systems.[2] Put more bluntly, when policymakers start legislating for Merlin instead of Microsoft, the result is frameworks that regulate fantasy, not function.

This misdirection was exemplified in 2016, when the European Parliament proposed considering "electronic personhood" for advanced autonomous robots.[3] While largely symbolic and met with skepticism from legal scholars, such discussions stem directly from the misperception of AI as autonomous entities capable of bearing rights and responsibilities.[4]

The proposal, drafted by Member of European Parliament Mady Delvaux, suggested creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause.[5] The framework would have required robot insurance and established legal liability mechanisms treating AI systems as quasi-legal entities.

The focus on theoretical personhood diverts attention from urgent, practical concerns about algorithmic accountability. While policymakers debate the hypothetical rights of non-existent sentient machines, real AI systems make consequential decisions about employment, criminal justice, financial services, and healthcare with limited transparency or recourse mechanisms. Algorithmic bias in hiring systems, opaque scoring mechanisms in lending, and unexplainable outcomes in criminal sentencing do not require consciousness to cause harm. For example, algorithmic hiring systems now screen millions of job applications annually and are known, in

some instances, to exhibit racial and gender[6] bias, systematically disadvantaging qualified candidates based on protected characteristics.[7]

Financial services provide another compelling example. Credit scoring algorithms make decisions that affect millions of consumers' access to housing, transportation, and economic opportunities. These systems often incorporate proxy variables that effectively discriminate against protected classes,[8] yet regulatory frameworks are only slowly shifting their attention from traditional lending practices to the already pervasive algorithmic decision-making processes.

This mystification also hobbles international regulatory coordination. When different jurisdictions operate under different mental and cultural models of what AI systems are, achieving a shared understanding of what exactly is being regulated becomes extremely difficult. The EU's early drafts of the AI Act, China's focus on controlling AI "behavior," and U.S. congressional hearings, which often focused on sci-fi scenarios, all reflect this deep confusion. This fragmented framing prevents coordinated responses to global challenges, as nations struggle to agree on whether they are regulating tools, agents, or something entirely new.

Misallocating Resources

The misallocation of attention is matched by a massive misallocation of resources, driven by the Heaven myth's promise of a "silver bullet." The embrace of techno-solutionism leads to flawed policies, wasted resources, and ultimately disappointment. When technology is sold as a panacea, policymakers pour money into

shiny new tools while putting aside the harder, messier work of tackling underlying social, economic, or political causes. The result is resource displacement: shiny tools take the stage while structural causes remain off-script. The resulting policies often fail not simply because the technology was oversold, but because the problems were misunderstood as technical rather than political and societal challenges.

This approach is especially attractive when real solutions demand difficult political conversations about resource allocation, power distribution, or systemic reform.[9] Consider how technology is often promoted as a solution to educational inequality. Rather than addressing disparities in school funding, teacher quality, or socioeconomic conditions that affect learning, policymakers invest in digital devices and software that promise to personalize learning and improve outcomes.[10]

As we have seen in Chapter 3, while educational technology can certainly provide value, it cannot overcome the inequalities that drive educational disparities. Students in under-resourced schools may receive tablets and learning software, but they still face challenges such as food insecurity, inadequate facilities, and inexperienced teachers that technology cannot address. The focus on technological solutions thus allows policymakers to appear responsive to educational challenges while avoiding more costly and contentious reforms to educational funding and social policy.

The pattern repeats across policy domains. In healthcare, electronic health records and AI diagnostic tools

promise to improve care quality and reduce costs, but they do not address underlying issues such as provider shortages, insurance coverage gaps, or social determinants of health. In criminal justice, predictive policing and risk assessment algorithms promise to reduce crime and improve fairness without addressing poverty, inequality, or systemic racism that drive criminal behavior.

Environmental policy also reveals this pattern. Geoengineering garners headlines with visions of planetary-scale interventions, ranging from carbon capture to solar radiation management and atmospheric modification, that would enable societies to continue their current patterns of fossil fuel consumption while mitigating their environmental effects. Yet many of these proposals function more as delaying tactics than solutions, postponing hard choices about carbon reduction.

These tech fixes are seductive not only because they appear bold but because they deflect blame. It is easier to believe in algorithmic answers than to confront decades of unjust policies and institutional failure. The focus on technological solutions also creates path dependency, making it more difficult to pursue alternative approaches when technological interventions prove inadequate. The displacement effect creates a self-reinforcing cycle, where vendors, administrators, and technical experts with vested interests in these new tools continue to promote them, justifying their expansion rather than questioning their effectiveness.[11]

Feeding Hype, Starving Reality

This leads to another devastating cost of hype: systematic resource misallocation that undermines long-term innovation and social benefit. When madness takes hold, common sense is the first casualty, swiftly followed by a substantial amount of money and a loss of democratic composure. Money flows toward anything with the right buzzwords, regardless of its actual utility or viability.

The AI investment boom since 2023is an example of this fever. Venture-capital funding for AI startups reached record levels, but much of the investment went to companies with questionable business models or marginal technological advantages.[12] Startups with minimal revenue but AI-powered pitches secured valuations that would have been unimaginable just a few years earlier. For example, in 2021, Character.AI raised $150 million, primarily based on the popularity of its AI chatbot platform.[13] The company's technology was interesting, but its business model relied primarily on subscription revenues from users who wanted to chat with AI personalities. When the novelty wore off and user engagement declined, the company struggled to justify its $1 billion valuation. The eventual outcome was a licensing deal with Google in 2024 that left investors with little to show for their enthusiasm.[14]

More troubling is the opportunity cost. The investment frenzy around AI crowded out funding for less glamorous but potentially more vital fields. While AI companies raised billions, research into renewable energy

storage, sustainable agriculture, and public health technologies struggled for attention.

Anthropic's $4 billion funding round from Amazon in 2024 illustrates the scale of this misallocation.[15] While the company's research into AI safety was valuable, the investment amount exceeded the total annual research budgets of many universities and public research institutions. The concentration of resources in a single company, regardless of its merits, raises questions about whether such investments are the best way to advance human knowledge.

The misallocation extends to people. Tech giants competed aggressively for AI researchers, offering enormous salaries and equity packages that drew talent away from academia and non-profit organizations. The "brain drain" was especially damaging to fields that are supposed to act as a check on corporate excess, such as AI safety and ethics.[16] This dynamic was highlighted when leading pioneers from institutions, including the University of Toronto, a key hub of AI innovation, departed for lucrative positions at companies such as Google and Uber, leaving academic labs struggling to retain senior talent.[17]

The Great Distraction: Picking Your Apocalypse

This massive distraction and misallocation are often justified by a "false choice" narrative, born from the Madness myth. We are told, in essence, to pick our apocalypse: die by a thousand biased cuts today, or be wiped out by a rogue superintelligence tomorrow. This is a dangerous misreading. Worrying about long-term risks and near-term harms is not a zero-sum game.[18] In

fact, the path to catastrophe is paved with the unaddressed failures of the present. The seemingly "small" problems—like algorithmic bias or AI generated misinformation—are not distractions from a "big" problem like AGI; they are the training grounds for it. Every time a company is not held accountable for a biased algorithm, it normalizes the idea that automated systems are exempt from human oversight. Every time misinformation is allowed to poison democratic discourse, it weakens the institutions needed to govern a more powerful technology.

Think of it like climate change. The immediate harms of pollution, like smog in our cities and plastic in our oceans, are not separate from the long-term threat of a collapsing biosphere. They are symptoms of the same systemic failure, two ends of a single, catastrophic spectrum. Ignoring the smog today doesn't provide more resources to fight global warming tomorrow; it just ensures the air is unbreathable long before the seas rise. So it is with AI.

The false choice serves the architects of madness perfectly. For the techno-optimists, it allows them to dismiss present-day harms as trivial bumps on the road to a glorious, inevitable future. For the doomsayers, it allows them to ignore the patient work of fixing today's broken systems in favor of a more dramatic crusade against a hypothetical apocalypse.

Mature governance requires us to reject this false choice. It demands we see the connection between the algorithm that denies someone a loan today and the hypothetical superintelligence that could destabilize the

world tomorrow. They are not two separate monsters. They are the same beast, simply at different stages of its life. To tame it, we must start now, by addressing the harms we can see, not by waiting for the ones we can only imagine.

Governance Whiplash: The Cost of Panic

The second significant cost of myth-based policy is governance whiplash. Driven by the Madness myth, our institutions swing wildly between the fever of hype and the chill of panic. This irrationality is the enemy of good governance, short-circuiting the careful and often slow processes of democratic deliberation. When a technology is framed as an immediate existential threat, the pressure to "do something"—anything—overwhelms the need to do the right thing.

This panic-driven process is highly visible in practice. As seen in Chapter 4, the legislative rush to address deepfakes is a good example. California's sixty-day election window ban and Texas's criminalization measures were crafted and passed without systematic hearings, expert testimony, or empirical assessment of the actual threat level. In Washington, D.C., congressional treatment of deepfakes as equivalent to weapons of mass destruction created an atmosphere where questioning the response became politically toxic, no matter what the evidence said. The same emergency logic played out with school bans on AI. New York City's prohibition of ChatGPT was a snap decision, made without any meaningful consultation with the teachers and students it would affect. The speed of implementation—in some

cases, within weeks of the tool's release—prevented the kind of thoughtful assessment that policy requires, typically.

The Global Contagion

Panic is highly contagious. In today's interconnected world, emotional responses in one jurisdiction quickly spread, creating global patterns of suboptimal governance.

Political scientists call this policy diffusion through panic, a process in which bad ideas spread across borders because the political cost of appearing unprepared is higher than the cost of a bad decision.[19]

The spillover effect is particularly damaging because it eliminates the opportunity for policy experimentation. When every country rushes to respond to the same inflated promise or exaggerated threat, the world loses the chance to learn from diverse approaches. The result is a form of global policy madness, where international coordination is based on shared delusions rather than shared evidence.

The Nirvana Fallacy

This whiplash between hype and panic leads to the Nirvana fallacy,[20] also known as the perfect solution fallacy, which pushes forward an unrealistic yet perfect-sounding solution instead of considering real-world solutions. The approach, born of either utopian promises or dystopian fears, demands perfection rather than improvement.

In a hype cycle, the fallacy rests on inflated promises. When AI companies claim to solve poverty, cure dis-

ease, or eliminate bias, they set expectations that no technology could ever meet. The inevitable disappointment leads to backlash that can kill funding and support for genuinely beneficial applications. The debate over AI education is the classic case. By demanding that AI tools pose no risk to academic integrity, officials overlooked the reality that concerns about "cheating" have accompanied every educational technology, from calculators[21] to computers.[22] Students have always found ways to cheat, with or without technology.

The pursuit of impossible standards doesn't just stifle innovation, it leads to rules disconnected from reality, built around exaggerated fears and imaginary futures.

Patterns of Flawed Regulation

When moral panic grips the policy machine, the result is not thoughtful governance but a stampede of reactive, ill-conceived rules. These frameworks satisfy the emotional demand to "do something," but they fail to address the root causes of problems and often create new harms. This panic-driven response typically falls into four predictable patterns.

#1. The Militarized Framing

When a technological challenge is framed as a national security threat, the policy response becomes militarized. Complex social and economic issues are flattened into us-versus-them conflicts, demanding decisive action, not democratic deliberation.[23] The cybersecurity debate is an example of this militarized thinking. When government officials invoke terms such as "cyber Pearl Harbor," they create a policy environment that justifies

extraordinary measures while marginalizing democratic oversight and the protection of civil liberties.[24]

The U.S. Department of Homeland Security's Cybersecurity and Infrastructure Security Agency, for instance, operates under authorities that would be controversial if applied to traditional law enforcement but become acceptable when framed as essential protection against foreign adversaries.[25] More generally, militarized framing enables the expansion of surveillance capabilities, the restriction of encryption technologies, and increased government control over digital infrastructure. History also shows that it inevitably leads to mission creep, where tools designed to counter foreign adversaries are turned inward, threatening civil liberties and political dissent.[26] This war-machine approach is often counterproductive. It sparks escalations with adversaries[27] and diverts resources from the defensive work of building secure systems, such as user education and institutional resilience, in favor of offensive capabilities that serve military functions.[28]

#2. The Binary Prohibition

The second pattern, born of pure panic, is regulation by prohibition. This approach is built on the deterministic fallacy that technology is inherently evil, ignoring the reality that its impact depends largely on how it is governed.[29] It treats technology as a binary choice between total prohibition and utter catastrophe, bypassing any nuanced discussion of risks and benefits.[30] The initial proposal for a temporary EU ban on facial recognition in public spaces, which was reported in January 2020,[31] illustrates this thinking. The idea was quickly dropped,

however,[32] and the AI Act opted for a tiered approach that bans specific harmful uses, such as real-time public biometric surveillance, while permitting other applications under strict safeguards.[33]

Similarly, calls for AI development moratoriums respond to existential risk concerns by proposing to halt all advanced AI research until safety issues can be completely resolved.[34] This response often backfires. When governments ban specific technologies rather than regulating their applications, they hinder the development of beneficial applications while doing little to deter malicious actors, who simply relocate to jurisdictions with weaker oversight.[35]

Prohibition also creates practical challenges. Enforcing such bans requires invasive monitoring, ironically expanding the surveillance capabilities the panic was meant to prevent.[36]

#3. The Symptom Regulation

The third pattern is the technocratic targeted approach. Here, a complex social problem attributed to an algorithm is addressed with a narrow technical fix, leaving the broader system intact. The algorithm is treated as a discrete entity that can be regulated independently, not as a component of larger arrangements.[37]

This is the logic behind relying heavily on automated content moderation to fight misinformation and harmful content, or on algorithmic audits to address discrimination. It does not address the underlying sources of misinformation production, nor does it improve media literacy education.[38] Similarly, while algorithmic

bias auditing requirements in hiring, lending, and criminal justice can help identify problematic patterns, they typically focus on algorithmic outputs rather than addressing underlying data biases or institutional discrimination that algorithmic systems reflect and amplify.[39]

Such interventions primarily target the symptom, rather than the disease. And they create collateral damage. Automated systems can censor legitimate speech while missing sophisticated propaganda.[40] Bias audits can be gamed by companies to create a veneer of compliance, without addressing substantive problems. Neither approach tackles the business models that incentivize the production of harmful content and discriminatory targeting practices.[41]

Worse, by creating an appearance of oversight (e.g., "this hiring algorithm has been audited"), this approach can legitimize the underlying problem, making it harder to challenge the injustice of the automated system.[42]

#4. The Moral Crusade

The final pattern is regulation as purification. This approach sees technology as a vector of moral contamination and seeks to purify the digital world through censorship and restriction, treating online content as inherently dangerous and citizens as perpetually vulnerable. Instead of empowering people with education that enables critical media consumption, the moral crusade opts for broad speech restrictions and age-gating requirements.[43]

Age-verification requirements for social media, for example, treat online interaction as inherently harmful to minors, ignoring the valuable social connections and educational resources these platforms can provide. A focus on restriction avoids looking into the more difficult work of addressing the underlying causes of youth mental health challenges, such as investing in better mental health resources, improving educational funding, or strengthening community programs.[44] Such systems raise significant concerns for the privacy and civil liberties of all users by requiring extensive identity documentation, yet they can be easily bypassed by tech-savvy youth.[45] The result is a system that compromises adult privacy while failing to protect its intended subjects, who may simply migrate to less regulated spaces.[46]

The moral crusade impulse also drives attempts to eliminate platform liability protections such as Section 230 of the U.S. Communications Decency Act, which shields online platforms from liability for most user-generated content and allows for good-faith content moderation. The argument is that making platforms legally responsible for user-generated content will compel them to remove harmful material. The more likely outcome, however, is a surge in defensive over-censorship that silences legitimate speech[47] and perversely provides a pretext for authoritarian censorship regimes.[48]

Ultimately, these flawed regulations, born of panic and hype, are often counterproductive. Content moderation requirements may increase censorship without reducing misinformation, algorithmic auditing may legit-

imize discrimination while creating compliance theater, and technology bans may prevent beneficial applications while failing to eliminate harmful uses.[49]

Evaded Accountability & Democratic Abdication

The final and perhaps greatest cost of myth-based governance is the erosion of democratic control. This happens in two distinct ways.

First, the Magic and Sin myths combine to create a sense of evaded accountability. By mystifying AI as an incomprehensible "black box" or scapegoating it as a "digital villain," we create an "accountability vacuum" where human actors can sidestep responsibility for the harms their systems cause.

Second, the Heaven myth creates democratic abdication. We are encouraged to defer to a priesthood of prophets—the tech CEOs and venture capitalists—who position themselves as visionaries qualified to steer our future. We abdicate our collective power out of faith in their promise of salvation.

The Accountability Vacuum: Blaming the Bots

This accountability vacuum arises when responsibility for algorithmic decisions becomes so diffuse that no one bears clear liability for the harmful outcomes.[50]

The Magic myth contributes by treating AI systems as agent-like entities. When AI is portrayed as possessing agency, it becomes easier to attribute negative outcomes to the code rather than to the human actors who designed, deployed, and profited from it. The Sin

myth completes the trick: the algorithm becomes a digital scapegoat, absorbing blame that should be directed at human choices about training data, optimization objectives, and deployment contexts. When AI systems cause harm, it's convenient to point to the algorithm or the bot as if they acted independently. This linguistic sleight of hand complicates efforts to hold developers, deployers, and decision-makers accountable for failures and biases.

The case of predictive policing algorithms illustrates this dynamic perfectly. When systems such as PredPol[51] and COMPAS[52] produce biased recommendations that perpetuate racial disparities, discussions often focus on "algorithmic bias" as if the bias emerged spontaneously from the technology. Such framing obscures the human decisions that created the outcomes: the choice to train systems on historically biased arrest data, the decision to prioritize arrest prediction over public safety, and the determination to deploy these systems despite their known limitations.

This diffusion of responsibility plagues other sectors. Healthcare AI is another example of accountability diffusion. When diagnostic algorithms produce incorrect recommendations that harm patients, responsibility becomes distributed among algorithm developers, healthcare institutions, and individual practitioners. The systems' complexity makes it difficult to trace specific decisions to specific actors, creating legal and ethical gray zones where patients have limited recourse against algorithmic harms. On social media, the problem extends to automated content moderation, where AI sys-

tems make millions of decisions daily about what constitutes acceptable speech. When these systems make controversial moderation decisions, platforms often describe the choices as algorithmic rather than what they are: policy decisions embedded in code that reflect specific human values and commercial interests.

This is not just a legal problem; it is a moral one. A society that deflects blame onto its tools abdicates its ethical responsibilities. To attribute harm to AI itself is to let its creators off the hook.

The Great Abdication: Deferring to Prophets

While the accountability vacuum allows creators to evade blame for past harms, democratic abdication enables them to preempt scrutiny of future ones. This occurs when our own democratic institutions amplify the prophets of the Heaven myth, treating corporate leaders as neutral arbiters of the public good.

In politics, we see this when lawmakers feature tech executives as star witnesses in AI hearings. In the media, this abdication occurs when journalists report industry predictions as objective news, amplifying corporate marketing without critical analysis. In academia, it's fueled by research funding structures that are heavily dependent on industry support and incentivize academic attitudes that support rather than question corporate salvation narratives.

This democratic abdication is perhaps the highest cost of all. It prevents the development of the institutional wisdom needed to govern, leaving us trapped in a cycle

of panic and reaction, unable to manage the very technologies we are told will save us.

A Critic's Toolkit: How to Spot the Myth Red Flags

Use these red flags and critical questions to challenge "policy hallucinations" and identify the four myths in headlines, product demos, and political speeches.

1. The MAGIC Myth

Red Flag: Hearing phrases like "It thinks . . ." "It understands . . ." "It feels . . ." "It's sentient . . ." or "It's a 'black box' that's too complex to explain."

The Distortion: This language is designed to hide human accountability by making a simple tool seem like a conscious, magical being.

The Reality Check: What is the specific **function**, not the *feeling*? Is it "thinking," or is it "statistically predicting the next word"? Who *built* it, and what were *their* goals?

2. The MADNESS Myth

Red Flag (Hype): "This is an arms race . . ." "We're falling behind . . ." "This will revolutionize everything . . ." "Trillion-dollar opportunity . . ."

Red Flag (Panic): "This is an existential risk . . ." "It will end humanity . . ." "We must ban it immediately . . ."

The Distortion: This emotional language (FOMO or fear) short-circuits democratic deliberation and forces a knee-jerk reaction.

The Reality Check: What is the **evidence**, not the *emotion*? Is this claim based on a *demonstrated capability* or a *speculative future*? Who benefits from me feeling this urgency?

3. The HEAVEN Myth

Red Flag: "This will solve . . ." (climate change, poverty, cancer). "This is inevitable progress . . ." "We're on a mission to benefit all humanity."

The Distortion: This positions tech leaders as prophets and asks for our faith, encouraging us to abdicate democratic oversight in favor of their "vision."

The Reality Check: What *human* problem, value, or trade-off is this "solution" ignoring? Who gets to define "progress"? And who *isn't* in the room to discuss the consequences?

4. The SIN Myth

Red Flag: "Job-killing robots . . ." "Addictive algorithms . . ." "Polarizing platforms . . ." (Any phrase that blames the *tool* for a *human* problem).

The Distortion: This creates a digital scapegoat to absorb blame for complex, systemic problems, letting human decision-makers and flawed policies off the hook.

The Reality Check: What *human choice* or *business model* is this narrative hiding? Is the problem the "job-killing robot," or is it a corporate policy of not re-investing in its workforce? Is the problem the "addictive algorithm," or is it a business model that *profits* from engagement?