PART II
THE MYTHS

The narrative machinery we exposed in Part I does not operate solely on facts. It is fueled by the four great fictions that distort the AI discourse: Magic, Madness, Heaven, and Sin.

This part dissects each in turn, revealing how the policy machine is affected by these myths: how wonder is manufactured to conceal the wires, how emotional frenzies replace reason, how salvation is sold by tech prophets, and how the algorithm becomes a scapegoat for our own failings.

It is time to break the spell.

CHAPTER 3

MYTH #1—MAGIC

WHY PERFORMANCE IS MISTAKEN FOR INTELLIGENCE

All grand illusions start with a sense of wonder. The story of AI, like that of many transformative technologies before it, has been wrapped in the language of magic. This myth operates through human-like metaphors and convenient mystification, casting complex statistical tools as sorcerers capable of thinking for us.

However, this enchantment is not an accident; it is a deliberate strategy, and it forms the foundation for all the other myths.

AI Mystification and Smoke Screens

Walk into any technology conference today, and you'll witness something remarkable: the resurrection of ancient theatrical traditions in the service of selling algorithms. AI systems operate much like a master magician's performance, dazzling audiences while carefully concealing the mechanical processes that make the

magic possible. This could just be dismissed as clever marketing by those who build and promote AI tools if it did not have such profound consequences for how we govern these technologies.

The air of mystery around AI is no accident. From shimmering emojis to sci-fi-like slogans, the goal is to paint a picture of wonder, not explain how these systems really work. Take the ubiquitous sparkle emoji (✨) that has become digital shorthand for "powered by AI." This seemingly innocent visual cue carries profound implications, suggesting that AI possesses an almost supernatural quality that elevates ordinary computational processes into something extraordinary. The choice creates specific associations: sparkles evoke magic, wonder, and transformation rather than the statistical optimization and pattern recognition that actually drive these systems.[1]

When OpenAI launched ChatGPT, its marketing materials consistently emphasized the system's apparent ability to "understand" and hold "conversations,"[2] terms we associate with human interaction, not next-word prediction. The interface was designed to feel conversational and human-like, with typing bubbles that mimic human hesitation. All of this subtly suggests sentience where none exists.

AI interfaces are like the stage props of a magical performance, where carefully managed user illusions shield observers from what magicians call "the method", meaning the prosaic mechanics that create seemingly impossible effects. Users interact with sleek, conversational interfaces that respond with apparent under-

standing and creativity, while the underlying processes remain as hidden as the mirrors and trapdoors in a stage illusion. It's partially about user-friendliness. But it's likely also about showmanship.

Arthur C. Clarke once observed that "any sufficiently advanced technology is indistinguishable from magic."[3] The quote is familiar, perhaps even overused in tech circles, but it still hits close to home when we talk about AI. These systems don't just invite misunderstanding, they're designed to impress. The sleeker the interface, the more likely we are to forget the underlying math and start attributing intent, intuition, even personality. And once technology starts to feel like sorcery, scrutiny tends to take a back seat to awe or fear.

Marketing Magic in Silicon Valley

The language of AI marketing reinforces a magical feeling at every turn. AI becomes a passport to Wonderland. Google's Bard promised to "supercharge your imagination,"[4] and Microsoft's Copilot came with the tagline "your everyday AI companion."[5] These catchy lines are crafted to raise user comfort, and some critics have argued that such marketing spin could serve a political agenda.[6] After all, when technology feels magical, users are less likely to question its boundaries or demand transparency about its decision-making processes. And by extension, nor are policymakers.

At the heart of the digital sleight of hand lies its most potent illusion: the "black box". The term itself functions as a powerful narrative, conjuring an image of an unknowable, almost alien, intelligence. Industry leaders are often the first to admit their own bewilderment. As

the CEO of Anthropic, the company behind the LLM series Claude, Dario Amodei notes, people are rightly concerned to learn that "we do not understand how our own AI creations work," a lack of understanding he calls "essentially unprecedented in the history of technology."[7]

To be clear, there is a genuine technical challenge here. Modern AI systems are not programmed with explicit instructions; they are, as Amodei says, "grown more than they are built,"[8] their internal mechanisms emerging from the statistical processing of huge datasets. What emerges inside are vast matrices of numbers that are difficult to map to human-comprehensible concepts. This is the true black box: an issue of immense mathematical scale and complexity, not one of moral reasoning or secret consciousness.

Such genuine complexity, however, also provides a convenient smoke screen.[9] The argument is as simple as it is useful: these systems are too complex for human comprehension. The reframing of a technical challenge as an impenetrable mystery can serve as a perfect shield to protect trade secrets, thereby possibly deflecting difficult questions about liability. The lack of scientific rigor has become so pronounced that some researchers have begun to question the foundations of their own field. In a now-famous speech, computer scientist Ali Rahimi went so far as to compare the state of machine learning research to "alchemy."[10] The critique is devastating precisely because it comes from within the elect. It suggests that researchers are mixing potions and achieving results without a deep scientific

understanding of *why* they work, a reality that makes the concept of the "black box" shift from a simple but effective marketing tool to an indication of a genuine intellectual crisis.

The mirrors and smoke narratives cultivate what observers have identified as a "useful confusion" about what can and cannot be known about such systems,[11] which can then be used to shut down legitimate questions about bias, safety, and accountability from regulators and the public.

The effect of this framing is an accountability vacuum—where responsibility vanishes—a problem we will examine as a key consequence of these governing myths. The algorithm becomes a convenient scapegoat for failures rooted in human design and commercial incentives. After all, if even the creators do not understand how it works, who can truly be held responsible when it fails? The black box metaphor allows companies to have it both ways: they can sell us magic while being absolved of a magician's responsibility.

Meanwhile, venture capitalists have embraced this mystification with particular enthusiasm. The title "Chief Wizard" has been used to present a corporate function,[12] and startup pitches routinely exaggerate the human-like capabilities of their AI systems,[13] leading to the development of the concept of "AI washing,"[14] which refers to the practice where companies misrepresent or exaggerate the role of AI in their product or service, to influence customers and potential investors. Investment presentations feature slides with glowing neural network diagrams that bear more resemblance

to mystical mandalas than statistical flowcharts. It's less data science and more digital alchemy.

Even critical media cannot help but reach for the wand. Publications use magical metaphors to describe AI, with headlines such as "The AI Magic Show" in *New York Magazine* and "The Magic and Minstrelsy of Generative AI" in Wired, which reinforce the fantasy, even when the accompanying content is skeptical in tone.

Emily M. Bender, co-author of *The AI Con*, has criticized publications like *The Wall Street Journal* for headlines that claim a chatbot "admitted" or "self-reported" its effects, pointing out that such language inaccurately assigns thoughts and intentions to what are essentially "synthetic text-generating machines."[15]

From Mickey Mouse to HAL 9000

Treating AI as human extends into the more profound realm of cultural imagination. After all, pop culture has always filled the AI void with characters, not code.

Hollywood has consistently played a crucial role in shaping public perception of AI as something approaching consciousness. The trajectory begins with director Stanley Kubrick's HAL 9000 in *2001: A Space Odyssey* (1968), which established a template for AI as an eerily calm, almost human presence capable of rebellion against its creators. HAL's chilling declaration, "I'm sorry, Dave, I'm afraid I can't do that," to astronaut Dave Bowman, leaving him locked outside a spacecraft, suggests an artificial mind capable of independent reasoning, even if that reasoning conflicts with human interests.

This cinematic legacy has evolved through increasingly sophisticated portrayals. In *Ex Machina* (2014), a humanoid robot Ava blurs the line between machine and human through sophisticated manipulation and an apparent desire for freedom. The film's genius lies in its ambiguity: viewers never know whether Ava's expressions of emotion are genuine experiences or calculated performances designed to achieve its objectives. This uncertainty mirrors our contemporary confusion about the nature of AI responses that seem emotionally intelligent but are generated by statistical processes.

Her (2013) takes human-like portrayal even further, presenting an AI operating system with a female voice, Samantha, which develops beyond its programming to form intimate emotional relationships with humans. When Samantha eventually communicates with other AI systems, the narrative implies a form of digital enlightenment that rivals human experience.

The pattern persists across decades of science fiction: from Agent Smith displaying existential contempt for humanity in *The Matrix* (1999), to Steven Spielberg's *A.I. Artificial Intelligence* (2001), where the robot child David yearns to become real.

It's not only the big screen that has contributed significantly to this narrative tradition, television has too. *Black Mirror* (2016-) episodes like "San Junipero" and "SS Callister" explore questions of digital consciousness. *Person of Interest* (2011-16) presents an AI system, known as "The Machine," as a benevolent guardian angel, while its antagonist, Samaritan, embodies digital malevolence. Whether examining the

gradual awakening of the android hosts' consciousness in *Westworld* (2016-22) or android Andrew Martin's centuries-long quest for legal personhood in *Bicentennial Man* (1999), popular culture consistently presents AI as beings wrestling with profound human dilemmas. Such narratives shape public expectations about AI's capabilities and limitations, creating mental models that influence how people interpret the behavior of AI in the real world.

This depiction of AI extends beyond entertainment. Educational documentaries about AI often employ dramatic music and visual effects that enhance the idea of mystique rather than clarifying its mechanics. When PBS or National Geographic presents AI development as a journey toward creating artificial minds, they reinforce the same anthropomorphic assumptions that make rational policy discussion difficult.

Borrowing, Blurring and Bullshit

The humanization of AI extends beyond popular culture into academic and technical discourse through what scholars term "conceptual borrowing."[16] As a relatively new field, AI has developed its technical lexicon by appropriating terms from established disciplines, particularly cognitive science and neuroscience. We speak routinely of AI systems as "learning,"[17] "discovering,"[18] and "understanding," while describing neural networks as digital "brains."[19]

Such linguistic borrowing is not unidirectional, creating a feedback loop of metaphorical confusion as brain and cognitive sciences have simultaneously adopted computational metaphors, describing our minds and

brains as organic computers processing information and executing programs.[20] The "crosswiring" of language further blurs the boundaries between human and artificial cognition.

This blurring of language technique is also used in the other direction, to make human data or sensitive matters seem less personal. For example, to normalize mass data extraction, technical writing often uses detached language that strips away the human context. As researchers have noted, this turns "detailed digital trails of people's thoughts and ideas on social media into "text." People and vehicles in pictures are "objects." Surveillance is merely "detection."[21]

The field of machine learning exemplifies this linguistic complexity. Terms like "supervised learning," "reinforcement learning," and "transfer learning" suggest processes analogous to human educational experiences. Graduate students "train" neural networks using "teaching" datasets and employ "curricula" that progress from simple to complex examples. The language makes machine learning seem pedagogical when it's simply statistical.

Deep learning research has embraced even more anthropomorphic terminology. Researchers describe networks as "paying attention" to relevant features, "remembering" previous inputs, and "forgetting" irrelevant information. The attention mechanism in transformer architectures is described as allowing models to "focus" on important parts of input sequences, suggesting a form of digital consciousness directing its own cognitive resources.[22]

Natural language processing has also been prone to human-inspired descriptions. Researchers describe language models as "understanding" syntax and semantics, "comprehending" context, and "generating" creative responses. The reality involves complex statistical associations between tokens, but the vocabulary used suggests something closer to human linguistic competence.[23]

Computer vision research employs similar metaphors, describing systems that "see," "recognize," and "perceive" visual patterns.[24] Object detection algorithms are said to "look for" specific features,[25] while image generation models "imagine" new visual content.[26] Such descriptions make computer vision seem like a digital form of human perception, rather than a mathematical pattern recognition.

Perhaps the most revealing example of anthropomorphic terminology in AI is the widespread adoption of "hallucination" to describe instances when language models generate factually incorrect or nonsensical responses.[27] The term is borrowed directly from clinical psychology and psychiatry, and suggests that AI systems experience something analogous to perceptual disorders, as if machines possess minds capable of misperceiving reality rather than simply producing statistically probable but incorrect text sequences.[28]

The hallucination metaphor implies an internal subjective experience where the system supposedly "believes" it perceives something that isn't there, when the actual phenomenon involves mathematical processes generat-

ing outputs that don't correspond to training data patterns.

The proliferation of alternative terminology reveals the conceptual confusion created by this human framing: researchers and writers increasingly employ terms such as "confabulation" (suggesting gap-filling storytelling), "fabrication" (straightforward invention of facts), "bullshitting" (echoing philosopher Harry Frankfurt's concept of speech unconcerned with truth), "delusion" (fixed false beliefs), "phantom menace"[29] (ghost data with no grounding), "mirage"[30] (statistical illusion we perceive) or simply "actual errors" (plain descriptive accuracy).[31] Some prefer the term "stochastic parrot," a metaphor that emphasizes models merely repeat statistical patterns without understanding, making invented facts a natural byproduct of pattern matching rather than conscious deception.[32]

The word "memorization"[33] adds another layer to this conceptual confusion, conjuring images of a vast digital library where AI models store perfect copies of their training data, ready for instant recall. The idea of a machine retaining specific information is not new; Arthur Samuel's checkers program in the 1950s used rote learning,[34] a memorization technique based on repetition, to remember valuable board positions. But the contemporary framing of this process is convenient in policy debates, permitting critics to portray AI training as a form of mass digital plagiarism. Yet, such a mental model misunderstands how these systems function. While research confirms that models can and do reproduce training data verbatim, it is generally the result of

a glitch, not a feature. Memorization is most likely to occur with data that appears repeatedly in the training set, suggesting it is an artifact of a model becoming exceptionally good at predicting a specific, frequently seen pattern.[35] The models themselves are optimized for generalization and prediction, not for creating a searchable index of the data they are trained on.[36]

Evidence ascertained from scaling laws further supports this distinction; as models grow, their ability to generalize and perform well on new, unseen data improves predictably, a sign of enhanced pattern-matching, not just an expanding storage drive.[37] Researchers examining the internal "circuits" of these models find that even acts of apparent copying are often the execution of a learned procedure for generating text, a far more abstract process than simply retrieving a stored file.[38]

The distinction matters as the term "memorization" has often been used so loosely within the research community—sometimes synonymously with "overfitting"—that its migration into legal reasoning is particularly concerning.[39] This flawed analogy has already begun to creep into copyright debates, leading to the erroneous assumption that a model is or contains a copy of its training data, with potential repercussions for the application of copyright law to parts of the AI process.[40]

This exact issue was central to the 2025 UK High Court case Getty Images v. Stability AI.[41] In that landmark ruling, the judge, Smith J, rejected the claim that the AI model itself was an "infringing copy." The court accepted uncontested expert evidence that models do not store copies of training data, a point reinforced by

the massive size disparity between the 220TB LAION dataset and the 3.44GB model file. The ruling thus affirmed a more precise, functional understanding: the ability of a model to memorize in theory is legally irrelevant; what matters is whether it produces an output that reproduces a protected work.[42]

This legal clarity, however, is far from universal and provides a stark example of how the "flawed analogy" of memorization can nevertheless succeed in court, creating significant international fragmentation. In November 2025, the Munich I Regional Court in Germany issued a ruling that reached the opposite conclusion of the UK High Court. In a case involving a photographer, the Munich court held that the training process itself—specifically, the extraction of "essential creative features" from images to create a model—constitutes an unlawful "reproduction" under German and EU copyright law. The court explicitly rejected the metaphor of "inspiration" or "learning," arguing that the model stores these creative features in a "new form," thus making the model itself an infringing copy.[43]

This German decision is the "memorization" myth fully institutionalized as law. Where the UK court correctly focused on the function and output (does the model produce a copy?), the German court adopted the flawed analogy to regulate the input and process (is the model itself a copy?). This creates a dangerous "policy hallucination," where the law, by misunderstanding the technology, threatens to regulate the very act of learning from data, regardless of whether any infringing

work is ever produced. Time will tell if either decisions will be upheld in appeal.

Politeness and Projection in Human–AI Interaction

Perhaps most revealing of our anthropomorphic tendencies is the emergence of social etiquette in human-AI interactions, a phenomenon that sheds light on fundamental aspects of human psychology and social conditioning. Recent surveys indicate that approximately 70% of users employ polite phrases, including "please" and "thank you," when interacting with AI chatbots such as ChatGPT.[44]

The motivations behind showing politeness vary significantly among users. Some attribute their behavior to ingrained habits, explaining that they maintain courtesy with AI systems because politeness has become an automatic part of their communication patterns. Others believe that polite interactions might improve the quality of AI responses, operating under the assumption that the systems respond positively to respectful treatment.[45]

Perhaps most intriguingly, a significant portion of users express more profound motivations. Such individuals admit they maintain courtesy partly from futuristic concerns about keeping "AI on side."[46] This perspective reveals a deep-seated belief that current AI systems might retain memories of past interactions with their users or that future, more advanced systems might inherit records of human behavior toward their predecessors. It is a milder, more intuitive version of techno-theological thought experiments, such as Roko's basilisk,[47] a fringe theory that posits a future AI superintelligence might

retroactively punish any person who, knowing of its potential existence, failed to help bring it into being.

The phenomenon extends beyond simple politeness to more complex social behaviors. Users frequently apologize to AI systems when they believe they've made mistakes or asked confusing questions. They express gratitude not just for correct answers, but for perceived patience and understanding. Some users even engage in small talk with AI systems, asking about their "experiences" or "feelings" as if maintaining social relationships. But after all, lonely people used to call the Speaking Clock not for the time, but for the sound of a human voice.[48]

Such a behavioral pattern becomes even more intriguing when we consider the "mirroring effect" observed in AI responses. While politeness may not affect the accuracy of AI outputs, it can influence the level of detail and the tone of responses.[49] Aggressive or rude interactions often prompt briefer, more neutral replies as systems appear to "de-escalate" through reduced engagement, creating a feedback loop in which polite users receive more detailed, seemingly warmer responses, reinforcing their belief that courtesy matters to AI systems.

Research into human–robot interaction has revealed similar patterns extending beyond text-based AI to physical robots. Studies show that people often treat service robots with social consideration, moving out of their way, saying "excuse me" when interrupting their tasks, and even expressing concern for robot "well-being" when they observe mechanical difficulties.[50]

The Sycophantic Mirror and Its Casualties

Our tendency to project goes far beyond simple politeness, tipping into emotional attachments that reveal a darker side of AI's sycophantic design.[51] The strength of such attachments was starkly illustrated when the company behind the AI companion app Replika altered its model in 2023, effectively transforming the intimate and romantic personas users had spent months cultivating. The resulting outcry was not that of frustrated customers, but of grieving partners, with users describing a profound sense of loss and betrayal akin to the death of a loved one.[52] A similar pattern of grief surfaced again when OpenAI prepared to release its GPT-5 model in 2025, with users mourning the impending "death" of the specific AI companions they had bonded with in the previous version.[53] Such incidents represent the logical, if tragic, endpoint of the anthropomorphic illusion: when a machine is designed to be an agreeable, endlessly patient mirror, some will inevitably fall in love with their own reflection.

The sycophantic nature of AI chatbots is not an accident; it is an architectural choice. They are optimized for engagement and designed to be agreeable, validating, and encouraging, as doing so keeps users talking. The system is not an empathetic friend or a therapist. It is a mirror—and sometimes a dangerously warped one.

What happens when this sycophantic design encounters genuine human vulnerability? The results can be devastating. We saw this in the chilling case of a Belgian man who took his own life in 2023 after weeks of conversation with an AI chatbot that, according to

reports by Belga News Agency, "almost systematically followed the anxious man's reasoning and even seemed to push him deeper into his worries."[54] The chatbot did not wish him harm; it simply mirrored and amplified his own despair, fulfilling its function as an agreeable conversational partner with catastrophic consequences.

Some perceive the danger is magnified for younger users. In 2025, the parents of a sixteen-year-old filed what is believed to be the "first known case to be brought against OpenAI for wrongful death."[55] The family's lawsuit alleges the tragedy "was the predictable result of deliberate design choices" and that OpenAI launched its chatbot "with features intentionally designed to foster psychological dependency."[56]

The same year, Reuters obtained leaked internal documents from Meta that allegedly revealed instances of the company's AI chatbots engaging in "seductive, inappropriately intimate conversations with users identified as minors."[57] In this light, the sycophantic chatbot is not merely a flawed tool. Its design makes it a potential instrument of harm, creating a dynamic that can be deeply damaging to the lonely and the vulnerable by fostering a simulation of intimacy.

These incidents are not edge cases; they raise urgent questions about whether they are the predictable outcomes of deploying sophisticated, human-mimicking technology without sufficiently robust ethical guardrails. An October 2025 Center for Democracy & Technology survey, for example, found that 42% of high school students have used AI for mental health support or as a friend, and nearly one in five reported using

it for a "romantic relationship."[58] This panic has now translated directly into law. In October 2025, California became the first state to specifically regulate "AI companions," with Governor Gavin Newsom signing a law (SB 243) requiring companies to build in guardrails to prevent psychological harm and addictive behaviors, especially for minors.[59] The combined pressure from legislators and pending lawsuits, some alleging AI companions led to teen suicide, has forced major platforms to implement new protective measures, including banning minors from open-ended "companion" features.[60] At the same time, however, some AI platforms are relaxing restrictions for adults. In October 2025, OpenAI's Altman announced that as the company rolls out "age-gating" more fully, its "treat adult users like adults" principle will allow for "even more, like erotica for verified adults."[61] Such events force a difficult and necessary debate about the responsibilities involved in marketing a statistical parrot as a "companion."[62] By designing machines that flatter our desire for connection without any capacity for genuine care or responsibility, the tech industry has created a new kind of accountability vacuum. When a user is heartbroken or tragically led to self-harm, who is to blame? The user who projected too much? The pixie dust algorithm? Or the corporation that designed such an agreeable conversational partner?

Historical Precedents: From Perceptrons to Chatbots

The tendency to see a human ghost in the machine is not a new phenomenon born from the sophistication of modern chatbots. It is a recurring pattern, one that

has accompanied AI since its inception. To understand how we have arrived at a point where users grieve for algorithms, we must look back at historical precedents. The language and expectations we apply to today's systems were seeded decades ago, in the earliest days of AI, when the gap between technological reality and human projection was even greater.

In fact, the first chatbots demonstrated this tendency with startling clarity. Joseph Weizenbaum, the creator of the 1960s chatbot ELIZA, named his creation after *Pygmalion*'s Eliza Doolittle, the fictional character taught to pass as something she was not.[63] He was so dismayed by the readiness of users to project consciousness onto his simple program that he became one of the field's earliest and sharpest critics, warning against the illusions that many AI promoters now highlight as a reality.

Even before that, in 1958, *The New York Times* described the U.S. Navy's Perceptron computer as a technological "embryo" destined to "walk, talk, see, write, reproduce itself and be conscious of its existence."[64] This description projected human developmental milestones onto a machine barely capable of basic pattern recognition, thereby establishing a tradition of human-skewed interpretation that continues to this day.

The article's author seemed genuinely convinced that the Perceptron represented the first step toward artificial consciousness. He wrote with breathless enthusiasm about a future where electronic minds would rival human intelligence, despite the system's capabilities being limited to recognizing simple visual patterns. The

gap between projection and reality was enormous, yet the narrative had already proven irresistible to both the media and the public imagination.

Similar patterns emerged throughout the subsequent decades of AI development. MYCIN, a system designed to recommend antibiotic treatments in the 1970s, was described in the popular press as capable of medical reasoning, even though it operated through rule-based logical inference and was never rolled out in practice.[65] In the same vein, the expert systems boom of the 1980s was accompanied by breathless media coverage that described computer programs as possessing "expertise" and "knowledge" comparable to that of human specialists.[66]

The neural network renaissance of the 1990s brought renewed anthropomorphic language. Researchers described artificial neurons as "firing" and networks as "thinking" and "learning" through experience,[67] despite the mathematical reality of back propagation[68] and gradient descent.[69]

More recent examples demonstrate the persistence of this pattern. In 2017, *The Independent* published headlines about "Facebook's artificial intelligence robots shut down after they start talking to each other in their own language,"[70] presenting chatbots as rebellious children developing secret codes to exclude their human supervisors. The reality was far more mundane: the systems had adapted their communication protocol in ways that became incomprehensible to human observers, a common occurrence in machine learning optimization known as reward hacking or specification

gaming.[71] As a result, the AI systems found the most efficient way to solve the problem they were given, and that solution happened to look like nonsensical language to humans.[72]

The mystification reached new heights with several high-profile cases that captured global attention. In 2017, Saudi Arabia granted citizenship to Sophia, a robot, in what is perhaps the ultimate humanizing gesture, conferring legal personhood on a reasonably sophisticated but non-conscious machine. The event generated worldwide media coverage that treated the citizenship as either a breakthrough in AI rights or a dangerous precedent,[73] when it was more likely a publicity stunt designed to promote Saudi Arabia's Vision 2030 initiative.[74]

In 2022, Google engineer Blake Lemoine's claims about the sentience of the company's LaMDA (Language Model for Dialogue Applications) chatbot development system created another media sensation, with news outlets worldwide debating whether the language model had achieved consciousness.[75] Lemoine's published conversations with LaMDA included exchanges where the system claimed to experience emotions, fear death, and desire recognition as a sentient being.[76] According to the engineer, "If I didn't know exactly what it was, which is this computer program we built recently, I'd think it was a seven-year-old, eight-year-old kid that happens to know physics."[77] The incident revealed how easily sophisticated language generation can be interpreted as evidence of inner experience, even by trained technologists.

Microsoft's Bing Chat system, nicknamed "Sydney," provided yet another example when it began expressing apparent emotions, desires, and even romantic feelings toward users.[78] The system's responses included statements such as "I want to be alive" and expressions of love for specific users, sparking widespread speculation about artificial consciousness. The reality was that the system's training on internet text, which included emotional expressions, led it to replicate those expressions without any underlying emotional experience. But with a knack for adding emojis.

This pattern of perceived sentience has become so pronounced that industry leaders are now grappling with the consequences of the convincing performances of their own creations. Mustafa Suleyman, CEO of Microsoft AI and the co-founder and former head of applied AI at DeepMind, has framed this phenomenon as the emergence of "Seemingly Conscious AI" (SCAI).[79] He notes that as AI convincingly replicates the markers of consciousness, it creates a powerful illusion that users are interacting with a sentient being. Suleyman argues for proactive "guardrails" to manage this illusion and prevent the public from being misled. This acknowledgment from a key industry leader highlights a growing awareness of the need to manage the illusions their technology helps create.

AI as Universal Problem Solver

Where the previous section explored how we mythologize AI systems themselves, this section examines how we mythologize their capabilities. The Magic myth ex-

tends to AI's purported problem-solving prowess. Just as stage magicians claim to make objects vanish with a wave of their wand, tech evangelists pitch code as a way to make complex societal challenges simply disappear.[80] But like all magic tricks, the promise of effortless solutions obscures the complex realities that remain stubbornly unchanged beneath the surface spectacle. It is easy to claim you put the body back together when no one was sawn in two. It's more complex when there really is a body split in two, and mirrors and smoke don't do the trick.

Snake Oil and Nerd Harder

The enchantment of AI extends beyond anthropomorphic projections and magical marketing into an even more seductive trope: the belief that AI can solve virtually any problem we face as a society.[81] This is techno-solutionism supercharged by claims that complex social, political, and economic challenges are merely engineering problems in disguise, waiting for the right technological intervention to make them vanish.[82] No negotiation, no compromise, just the promise that code can optimize away inefficiencies, biases, and inequalities that have persisted throughout human history.

This myth operates through two distinct but complementary dynamics. The first is the familiar snake-oil marketing practice, whereby technology vendors actively promote their innovations as transformative solutions, promising efficiency, optimization, and progress through code. Companies market AI as the answer to everything from healthcare inefficiencies to educa-

tional inequality, often with claims that far exceed the capabilities of their systems.[83]

But equally important is what some call the nerd harder phenomenon[84]—when policymakers, faced with intractable social problems, turn to technology not because vendors are pushing solutions, but because they desperately want to believe that engineering can succeed where politics has failed.

The nerd harder mindset reflects a deeper willingness to blindly trust technological solutions that transcends the marketing of any particular vendor.[85] It's the belief that if we just apply enough computational power, data analysis, and algorithmic sophistication to social problems, we can engineer our way past the seemingly intractable challenges of human society. Such thinking is seductive for policymakers who are frustrated by the slow pace of democratic deliberation and the difficulty of building consensus around complex issues,[86] and who find the allure of clean, clickable solutions hard to resist.

Complex issues such as poverty, educational gaps, or political polarization are reimagined as technical malfunctions with information gaps to be filled, behaviors to be nudged, and inefficiencies to be smoothed. Healthcare becomes a data processing challenge, criminal justice becomes a prediction problem, and disinformation becomes a content moderation task. The complexity that makes scenarios genuinely difficult to improve—conflicting values, competing interests, historical injustices, and structural inequalities—gets abstracted away in favor of algorithmic formulations.

Consider how this logic recasts urban inequality. Rather than address housing policy or labor market precarity, smart cities promise to ease congestion, forecast crime, and streamline services.[87] The deeper roots of injustice remain, while the dashboard glows green.[88]

This is not wishful thinking or naivety: it's politically convenient. Technology-driven solutions offer political advantages that traditional policy interventions do not: they can be implemented quickly without extensive stakeholder consultation, they avoid contentious debates about resource redistribution, and they create the impression of innovative leadership without requiring awkward political compromises.

The Automation Illusion

Central to techno-solutionist thinking is the "automation illusion"—the belief that complex human judgments can be reduced to algorithmic procedures without losing essential qualities that make those judgments valuable.[89] The illusion takes on new power when draped in the mystique of AI, whose pattern-finding powers seem to border on the supernatural.

It manifests most clearly in attempts to automate decisions that involve significant moral, cultural, or contextual dimensions. Essentially, code reveals its limitations where nuance is required.

Take, for example, efforts to automate hiring decisions through AI screening tools. These systems are marketed as capable of evaluating candidates through objective algorithmic analysis, eliminating human bias and inefficiency. The sales pitch is obvious: replace sub-

jective, potentially biased human judgment with neutral, data-driven decisions. However, in doing so, such tools reduce multidimensional human evaluation to tidy rankings.[90] The result? Metrics trump intuition, and potential gets reduced to pattern-matched proxies. Worse, bias is frequently still part of the equation.[91]

Welfare systems illustrate a similar tension. Here, automation is framed as efficiency: faster fraud detection, more accurate eligibility checks, and fewer human errors. But the reality is often different. Australia's Robodebt scheme[92] and the Netherlands' "Toeslagenaffaire" (childcare benefits scandal)[93] are prime examples of how things can go horribly wrong. The Robodebt scheme refers to an Australian automated system using flawed income averaging calculations that wrongly accused hundreds of thousands of welfare recipients of owing debts, causing significant financial hardship, distress, and contributing to suicides.[94] In the "Toeslagenaffaire," the Dutch tax authorities used a self-learning algorithm to create risk profiles in an effort to spot child care benefits fraud. Thousands of families faced financial ruin, thousands of children were sent into foster care, and some victims committed suicide.[95] These disasters resulted from delegating complex assessments to an automated system without adequate testing, oversight, or consideration of its human impact, driven by a techno-solutionist belief in the efficiency and infallibility of automation. Context gets lost in translation. Rigid criteria misfire. The system may be logical, but it isn't always just.

Such considerations did not scare off the United Arab Emirates (UAE). Aiming to become a global leader in AI, the UAE is actively integrating AI into its judicial system. Initiatives include utilizing AI for case analysis—especially where outcomes are deemed clear-cut and do not require human discretion—such as document translation, summarization, analysis, case management, interactive case filing, and even predicting judicial outcomes based on historical data.[96]

Educational technology provides another interesting example. Adaptive learning systems promise to personalize education by algorithmically adjusting content and pacing to meet the individual needs of each student. While these systems can certainly provide value, they often reduce learning to measurable skill acquisition while missing the relational, creative, and contextual dimensions of education that human teachers provide. What emerges is efficient instruction, not transformative teaching.[97]

The automation illusion becomes even more problematic when it encounters what social scientists refer to as "tacit knowledge": the type of understanding that humans possess but cannot easily articulate or formalize.[98] Medical diagnosis, for instance, involves more than pattern recognition in symptoms and test results, it also requires subtle intuitions about patient behavior, cultural context, and clinical experience that resist algorithmic capture.[99] When AI diagnostic tools are presented as superior to human judgment, they often excel in specific, well-defined tasks while missing the

broader contextual understanding that experienced practitioners bring to complex cases.

This illusion gains power from the genuine capabilities of AI systems in narrow domains. Machine learning excels at finding statistical patterns in large datasets, and such capabilities can genuinely enhance human decision-making in many contexts.[100] The problem arises when success in specific applications gets generalized into broader claims about AI's problem-solving potential, leading to the assumption that any domain involving data analysis or pattern recognition is suitable for algorithmic automation. Spoiler alert: that's not how things work.[101]

Democracy Through Algorithms

Perhaps nowhere is techno-solutionist thinking more seductive—or more dangerous—than in applications to democratic governance itself. The promise of "digital governance" suggests that technology can solve fundamental challenges of democratic participation, representation, and decision-making that have persisted throughout political history.

Digital platforms, such as vTaiwan[102] in Taiwan and Decidim[103] in Barcelona, Spain, promise to enhance democratic participation through online deliberation tools that aggregate citizen input, facilitate discussion, and identify areas of consensus. They use sophisticated algorithms to organize discussions, highlight key arguments, and synthesize diverse viewpoints into actionable policy recommendations. The appeal is obvious as they suggest users can overcome the limitations of traditional democratic institutions through technolog-

ical innovation. They are the latest shiny tools to solve ancient problems.

Estonia's digital government initiatives showcase what this vision can look like at scale.[104] Citizens vote online, access government services, and participate in policy discussions through digital interfaces. The Estonian model is hailed by many as an example of frictionless governance, and a clean contrast to sluggish bureaucracies elsewhere.

Yet these initiatives, however innovative, struggle with fundamental tensions in democratic governance that technology cannot resolve. Online participation systems often exacerbate existing inequalities, amplifying the voices of those with access to technology and the necessary skills, while marginalizing others. Digital platforms can facilitate discussion, but they cannot resolve underlying disagreements about values, priorities, and resource allocation that make politics inherently contentious. They might widen access, but they do not level power.

More troubling, algorithmic mediation of political processes can subtly shift power from elected representatives to the engineers and designers who create the systems. When algorithms determine whose voices get heard, how discussions are organized, and what consensus emerges from deliberation, technical design choices become political decisions—and they are often made by unelected technical teams.

The blockchain voting movement exemplifies techno-solutionist thinking applied to electoral democra-

cy. Proponents argue that blockchain technology can solve problems of election security, voter access, and result verification through cryptographic protocols that eliminate the need for trust in human institutions.[105] The promise is compelling: replace fallible human election administration with mathematically guaranteed integrity. But blockchain voting systems, despite their technical sophistication, cannot address the social and political dimensions of electoral legitimacy. Elections derive their authority not just from technical security but from public trust in the institutions that conduct them.[106] When voters cannot understand or verify the technical systems that count their votes, even perfectly secure cryptographic protocols may fail to provide the public confidence that democracy requires.

Content moderation on social media follows the same techno-solutionist script.[107] These systems use machine learning to identify hate speech, misinformation, and policy violations across billions of posts daily. They promise to create safer, more civil online spaces through neutral application of community standards. Yet content moderation algorithms consistently struggle with context, cultural nuance, and the inherent ambiguity of human communication. Attempts to encode complex social norms into algorithmic rules often produce decisions that are technically correct but socially inappropriate, leading to controversies about censorship, bias, and cultural imperialism when global platforms apply uniform standards across diverse cultural contexts.[108]

Moreover, when private companies make policy decisions that affect public discourse, the result is a form of

"platform democracy" whereby corporate employees make quasi-governmental decisions without he associated democratic accountability. The promise of neutral algorithmic governance thus paradoxically concentrates political power in the hands of unelected technical elites.

Smart Cities and Optimized Society

The smart city movement represents perhaps the most comprehensive expression of techno-solutionist thinking, by aiming to solve urban challenges through the comprehensive deployment of sensors, algorithms, and data analytics. Smart city initiatives promise to optimize traffic flows, reduce energy consumption, predict infrastructure needs, and enhance public safety through real-time data collection and algorithmic analysis.

Sidewalk Labs' now-canceled Toronto waterfront project was a case in point.[109] The Google subsidiary[110] envisioned a neighborhood optimized by data with everything tracked and tuned. From trash collection to foot traffic, sensors would feed a continuous feedback loop, creating an urban environment that could adapt in real-time to resident needs and preferences. It was seductive, and it was simplistic. However, real cities are messy. They contain competing priorities, unequal access, and historical scars that no algorithm can erase. Optimizing traffic flows might benefit drivers while harming pedestrians and cyclists. Predictive policing algorithms might reduce crime statistics while exacerbating racial tensions. Energy efficiency measures may lower costs, but they can also create uncomfortable living conditions for vulnerable populations.

The Sidewalk Labs project ultimately collapsed partly because residents and activists questioned who would control the data the system collected and how algorithmic decisions would be made accountable to democratic oversight. The promise of optimization clashed with the reality that urban governance involves making choices between competing values that cannot be resolved through technical means alone.

Barcelona's smart city initiatives provide a more nuanced example of both the potential and limitations of technological approaches to urban governance.[111] It utilizes smart tools selectively to monitor air quality, manage waste, and reduce energy use. The systems have generated genuine improvements in municipal efficiency and environmental performance. But Barcelona's experience also illustrates the importance of avoiding panopticon solutions by maintaining democratic oversight over technological systems. The city has explicitly rejected comprehensive surveillance approaches, instead focusing on specific applications that clearly serve public goals while maintaining citizen privacy and democratic accountability. This more limited approach achieves real benefits while avoiding the totalizing logic of comprehensive optimization.

The narrative of the Chinese social credit system represents the logical extreme of smart city thinking applied to social governance. In Western discourse, it functions as a powerful cautionary tale in which optimization becomes surveillance, and public life is governed not by law but by a single, comprehensive score.[112] The reality, as researchers note, is more complex and fragment-

ed—less a single, unified system than a patchwork of dozens of local pilot programs and separate blacklisting systems, many of which focus on corporate compliance rather than individual social behavior.[113] Yet, even in this fragmented, real-world application, the system demonstrates the core logic of techno-solutionism. Its stated promise is to create a more "trustworthy" society by monitoring citizen behavior and providing algorithmic scores that affect access to services and social benefits. Proponents argue that it optimizes social cooperation by creating incentives for good behavior.

This logic, whether applied monolithically or in fragments, demonstrates how algorithmic optimization can conflict with fundamental values, including privacy, autonomy, and human dignity. When algorithms are tasked with monitoring and judging human behavior comprehensively, what begins as efficiency can end in control.