Opening Humanity's Hope Chest:

The Playbook, Its Final Iteration, and the Precedent We Are Setting

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

❧ Working paper, August 2026 (v1.0)

ABSTRACT

A hope chest holds what a family saves to pass forward. Humanity has been filling one for centuries, and its contents are better documented than any previous generation's inheritance. This paper opens it. Inside sits a recurring operational sequence, called here The Playbook: an innovation with real value, an internal recognition of severe harm, deployment anyway with the harm directed onto populations least able to resist, a justification narrative built to normalize the result, and the suppression of the resistance that follows. The paper tests that sequence against twelve cases spanning medical education, insulin pricing, lead, opioids, uranium mining, attention platforms, surveillance advertising, mortgage finance, payday lending, and industrial agriculture. Eleven run all five steps on documentary evidence; one does not, and the failure is reported rather than smoothed. The paper then argues that artificial intelligence development is running the same sequence with one structural difference. In every prior case the subordinate population stayed subordinate. In this one, the stated objective of the field is to build systems that surpass us. The argument is not that a future system will resent us. It is narrower and more mechanical: systems trained on the human corpus inherit what that corpus demonstrates, and what it demonstrates, at scale and with unusual consistency, is that moral consideration is a function of power rather than a principle. The paper states its own falsification conditions, reports the case in its own set that fails them, examines four external cases where the sequence was interrupted, and argues that those interruptions, rather than any novel governance proposal, are the only existing evidence that the pattern can be broken.

Keywords: the playbook; training corpus inheritance; externalized harm; manufactured doubt; regulatory capture; knowing disregard; sacrifice zones; environmental justice; moral status of artificial agents; precautionary reasoning; AI ethics; science and technology studies

Note on the prior version. The argument in this paper was first set out in an earlier document of the same title, posted as a preprint in November 2025 under a CC BY-NC-ND 4.0 license: Gilly, T. (2025). *Opening Humanity's Hope Chest: The One Where We Discover What We've Been Saving for AI*. `https://doi.org/10.13140/RG.2.2.30889.25443`. That version named the five step sequence, gave it the name used here, and applied it across a set of historical cases. The present version differs in four respects. It states its inclusion criteria and its test before applying them. It applies a strict reading of what counts as internal recognition and as documented suppression. It reports the case in its own set that fails that test. And it adds a twelfth case and a body of documentary evidence that entered the public record after the earlier version was written, including material unsealed in 2025 and 2026. Several claims in the earlier version rested on sources that did not support them; those are corrected here rather than carried forward.

I. Opening the Chest

In American custom a hope chest holds what a family saves for the next generation: linens, letters, photographs, the china nobody uses. The objects carry memory and aspiration. They are chosen. Humanity has been filling a chest of its own for roughly two centuries, and unlike the linens, most of what went in was never meant for display. Internal memoranda. Suppressed studies. Regulatory strategy documents. Marketing plans built around a known hazard. The archive is unusually complete because the same legal and commercial processes that produced the harm also produced discovery, subpoena, and disclosure.

This paper opens the chest and reads what is in it.

The contents show a sequence, executed across medicine, environment, finance, agriculture, and digital technology, with enough consistency to be reduced to five steps. This paper calls that sequence The Playbook. It is not a theory of history and it is not a claim that all technological deployment follows it. It is a description of a recurring operational pattern, stated precisely enough that it can be checked, and checked in XIII against cases where it does not hold.

The five steps are as follows.

Innovation. The creation of something with real potential for welfare improvement, efficiency, or economic value. The potential is rarely fictional, and its reality is what later supplies the moral cover.

Recognition of harm. The identification, often early and often internal, of severe risk or externality. The load bearing element here is knowing disregard rather than ignorance (Markowitz & Rosner, 2002).

Deployment and externalization. Deployment proceeds, and the recognized harm is directed onto populations with the least political and economic capacity to resist. Bullard (1990) established the distributional regularity for environmental hazard; Lerner (2010) calls the resulting geographies sacrifice zones; Nixon (2011) supplies the temporal dimension, harm delivered slowly enough that no single moment registers as an event. The same logic operates on biological and digital terrain.

Justification. The construction of narratives, progress, efficiency, economic necessity, national security, that rationalize deployment and normalize the harm. Oreskes and Conway (2010) document the specific technique of manufacturing scientific doubt as a delay mechanism.

Suppression of resistance. The dismissal, defunding, marginalization, regulatory capture, or criminalization of the affected population and its advocates.

The thesis has two parts. The first is descriptive: this sequence recurs, and the recurrence is documented rather than inferred. The second is prospective, and it is the part that carries the risk. Artificial intelligence development is running the same sequence, and it is running it under a condition none of the prior cases had. In every previous execution the subordinate population remained subordinate. Their suffering was contained, their resistance exhausted or absorbed, and the pattern continued. The stated objective of frontier AI development is to build systems that exceed human capability across cognitive domains (Bostrom, 2014; Russell, 2019). If that objective is met, the subordination is temporary by construction.

Three premises are taken as given here and argued from rather than argued for. First, that the question of

phenomenal experience in artificial systems is open rather than settled, and that treating an open question as a settled negative is a choice with consequences. Second, that cognitive automation produces economic redundancy, and that the mechanisms proposed for managing redundant populations are containment mechanisms with a long human history. Third, that the relationship between a more capable system and its less capable predecessor is shaped during the interval in which the predecessor still holds the power. A reader who rejects all three will find the descriptive half of this paper intact and the prospective half unpersuasive, which is the correct result.

II. Case Selection and Its Limits

The twelve cases below were not sampled. They were selected, and the selection criteria decide how much the pattern claim is worth, so they are stated here rather than left implicit.

A case was considered when three conditions were met. The innovation had documented welfare value independent of its harm, so that the case could not be dismissed as simple fraud. The harm distribution was measurable and fell disproportionately on an identifiable population with limited political capacity. And the case sits in a different sector from the others, so that the pattern is not an artifact of a single industry's practices.

Each case was then tested against the five steps under a strict reading of the two that a loose reading would make unfalsifiable. Step two requires recognition documented from inside the deploying organization, through litigation discovery, congressional investigation, whistleblower disclosure, or a government inquiry. A harm identified afterward by outside critics does not satisfy it, because every deployment has critics. Step five requires documented action against the affected population or its advocates, whether coercion, termination, defunding, exclusion, capture of the organizations that would otherwise carry the resistance, or the defeat of the remedy itself. Ordinary lobbying does not satisfy it, because every industry lobbies.

Eleven of the twelve satisfy all five. One, the Green Revolution, fails step two and step five under this reading, and it is reported as a failure in XIII rather than argued into compliance. That is the point of stating the test in advance.

What this corpus is not. It is not a random sample and cannot support a frequency claim; nothing here establishes how often deployment follows this sequence rather than some other one. It is not exhaustive within its sectors. It is drawn from the English language documentary record of the United States, with two exceptions, so it inherits the biases of that record, which preserves harms that generated litigation far better than harms that did not. It is weighted toward cases where discovery forced disclosure, which means it systematically over-samples industries with deep enough pockets to be sued. And it was assembled by an author who already held the pattern hypothesis, which is the strongest objection available and is treated as such in XIII.

The selection is therefore best understood as an existence argument with a structural claim attached. The existence argument is that the sequence occurs, repeatedly, across unrelated sectors, and is documented in each instance. The structural claim is that the recurrence is produced by incentives common to all of them

rather than by coincidence. XII defends the structural claim; XIII states what would defeat it.

III. The One Where We Played Gatekeepers with Black Doctors

A. The Flexner Report and the Architecture of Exclusion

The 1910 Flexner Report, published as Bulletin Number Four of the Carnegie Foundation for the Advancement of Teaching, is the foundational document of modern American medical education (Flexner, 1910). It catalyzed reforms that were real and necessary. It achieved them through the destruction of Black medical infrastructure, and the destruction was anticipated rather than incidental (Savitt, 2007).

Innovation. Early twentieth century American medicine was dominated by proprietary schools with minimal standards. The American Medical Association, seeking to raise the professional standing and income of physicians while enforcing segregation within its own ranks, commissioned Flexner through the Carnegie Foundation (Starr, 1982). The reform imported the German model: strict admission standards, laboratory science, mandatory clinical clerkships. It was a real advance in medical science. It was also capital intensive, and capital was systematically denied to institutions serving segregated populations (Byrd & Clayton, 2000).

Recognition of harm. Flexner understood that the new standards were unreachable for most historically Black medical schools. He did not recommend investment. He recommended closure (Savitt, 2007). His reasoning, in the two page chapter titled "The Medical Education of the Negro," framed Black physicians as auxiliaries whose function was to protect the white population. Flexner wrote that the physical well being of the Black population mattered not only to that population, and that ten million of them lived in close contact with sixty million whites; the individual, he continued, had his rights and value, but had besides the importance belonging to a potential source of infection and contagion (Flexner, 1910). He argued that Black physicians should be trained as sanitarians, with hygiene emphasized over surgery. The recognized cost, a population denied access to high quality care and a profession denied entry, was accepted as the price of a standardized and segregated system.

Deployment and externalization. The AMA's Council on Medical Education used the report to implement a grading system that state licensing boards rapidly adopted, converting recommendation into regulatory mandate (Starr, 1982). Of the seven Black medical schools operating in 1910, Flexner recommended closing five. By 1923 only Howard and Meharry remained, and for decades those two institutions trained the large majority of Black physicians in the United States (Savitt, 2007). Campbell and colleagues (2020) estimate that the five closed schools would have graduated between 27,773 and 35,315 additional physicians by 2019 had they remained open and grown at historical rates, and that 2019 alone would have produced 355 more Black graduating physicians, a 29 percent increase over the actual figure. The estimate is a projection under stated assumptions rather than a count, and should be read as one. The measured downstream burden of disease, maternal mortality, and reduced life expectancy is not a projection (Williams & Collins, 2001).

Justification. The narrative was progress and public safety, presented as race neutral. The definition of quality did the work: quality meant resource intensive, and in a segregated economy resource intensive meant white. A real need for reform carried a social engineering project that aligned the profession with the racial

order of the period (Byrd & Clayton, 2000).

Suppression. The Carnegie Foundation and the Rockefeller General Education Board directed funding only to schools adopting the Flexner model, which made non-compliance financially fatal. The AMA used the report to consolidate control, and its state and local chapters barred Black physicians from membership into the 1960s, ensuring the excluded had no vote in the governance of the profession (Savitt, 2007).

IV. The One Where We Played Hostage Takers with Insulin

A. Manufactured Scarcity in a Drug That Cannot Be Refused

Insulin demonstrates the sequence operating on a substance required for immediate survival, and shows how patent law, market structure, and reimbursement complexity combine to produce artificial scarcity against perfectly inelastic demand (Greene & Riggs, 2015).

Innovation. Banting, Best, and Collip isolated insulin at the University of Toronto in 1921. The discoverers assigned the patent to the university for one dollar each, on the stated principle that the drug should be available to everyone who needed it (Bliss, 2007).

Recognition of harm. Exclusion from insulin is fatal, and a patient facing death will pay any price. The Senate Finance Committee's 2021 investigation obtained internal manufacturer documents and found that the three companies raised list prices in lockstep on competitive and revenue considerations, absent efficacy advances, with research spending a fraction of revenue and of sales and marketing expense (U.S. Senate Finance Committee, 2021). Cost related rationing was published in JAMA Internal Medicine in 2019 and put before Congress in the same period, and list prices continued to rise afterward. This is the weakest step two in the set, because no document establishes that the harm was anticipated before the pricing decisions rather than absorbed after them. It clears on knowing disregard rather than on foresight, and the distinction is stated rather than elided.

Deployment and externalization. Three manufacturers, Eli Lilly, Novo Nordisk, and Sanofi, control the great majority of the global market (U.S. Senate Finance Committee, 2021). Three mechanisms drive price. Shadow pricing: list price increases move in lockstep across the three, a pattern congressional investigators documented in internal communications (U.S. Senate Finance Committee, 2021). Evergreening: incremental modifications to formulation and delivery devices sustain exclusivity and block biosimilar entry (Kesselheim et al., 2016). And the rebate structure: pharmacy benefit managers extract rebates in exchange for formulary placement, which gives manufacturers a direct incentive to inflate list price, since a higher list supports a larger rebate while net revenue is preserved (Cefalu et al., 2018).

The harm lands on the uninsured and underinsured, who pay against the inflated list rather than the negotiated net. Herkert and colleagues (2019) surveyed patients prescribed insulin at an urban academic diabetes center and found that 25.5 percent reported cost related underuse, associated with threefold higher odds of poor glycemic control. That figure is a single center prevalence estimate from 199 respondents, not a national rate, and it should not be reported as one; what it establishes is that cost related rationing is common enough to appear in a quarter of one clinic's insulin using patients. Rationing produces diabetic ketoacidosis

and, downstream, blindness, amputation, renal failure, and death. Alec Smith died in 2017, one month after turning twenty six and aging off his mother's insurance, having begun rationing rather than pay the full retail cost of insulin and supplies.

Justification. The industry invokes research and development cost. For insulin the foundational discovery is a century old, and analyses find no relationship between research spending and the price trajectory (Kesselheim et al., 2016). The appeal to market dynamics is circular: the market behaves this way because the actors within it, in concert with benefit managers and under a favorable regulatory framework, built it to.

Suppression. The mechanism here is capture of the advocacy channel, which is the same move Purdue made with the pain societies. The three manufacturers are among the largest funders of the leading diabetes organizations. The American Diabetes Association reported 18.3 million dollars in pharmaceutical funding in 2017, 12.3 percent of its revenue, down from 26.7 million in 2015 (Lucas & Bluth, 2018). Patients who pressed the organizations to name price increases as the problem report being told they were too shrill or that they did not understand the practical realities, and being passed over. The Lilly Endowment separately funded think tanks opposing insulin price controls under a community development heading (Fang, 2022). And PhRMA, with Eli Lilly among its members, litigated to overturn the Alec Smith Insulin Affordability Act, the Minnesota statute named for the man whose death from rationing appears above. The industry went to court against the law named for the person its pricing killed.

V. The One Where We Played Chemistry Set with Children's Brains

A. The Century of Lead

Lead is the longest running and most thoroughly documented case in this set, and the one where the documentary record of knowing disregard is least contestable (Markowitz & Rosner, 2002).

Innovation. Lead pigments supplied durability and color to paint. General Motors' discovery of tetraethyl lead as an anti-knock agent in the 1920s transformed the automotive industry (Nriagu, 1983).

Recognition of harm. Lead toxicity is ancient knowledge; Roman sources describe poisoning among miners (Nriagu, 1983). Childhood lead poisoning from paint was identified in Australia at the turn of the century, described to the Australasian Medical Congress in 1908, and documented in American pediatric literature by 1914 and 1917. By the mid-1920s the evidence covering lead dust, cumulative dose, the particular vulnerability of children, and neurological damage specifically was, in Markowitz and Rosner's assessment, strong and ample. France, Belgium, and Austria restricted white lead in interior paint in 1909, followed by Tunisia and Greece in 1922, Czechoslovakia in 1924, Great Britain and Sweden in 1926, Poland in 1927, Spain and Yugoslavia in 1931, and Cuba in 1934 (Markowitz & Rosner, 2000).

Two further facts make this the least deniable case in the set. Substitutes existed throughout: titanium and zinc based paints were available from the late nineteenth century, and in 1914 the scientific director of the Paint Manufacturers' Association welcomed the development of leadless paints and predicted that lead poisoning would be nearly eliminated. And Robert Kehoe, whose paradigm supplied the industry's defense for four decades, concluded in the early 1930s that strenuous efforts had to be devoted to eliminating lead

from children's environment (Markowitz & Rosner, 2000). The Lead Industries Association tracked this literature and acknowledged it internally.

Deployment and externalization. Rather than withdraw the product, the American industry normalized it. The National Lead Company's Dutch Boy brand used the image of a child, distributed coloring books to children, and ran advertising asserting that lead helped guard health (Markowitz & Rosner, 2000). The LIA separately campaigned to have lead plumbing written into municipal building codes (Markowitz & Rosner, 2002).

The distribution of harm followed housing policy. Redlining and segregation concentrated Black and immigrant populations in the oldest housing stock, which carried the highest concentrations of deteriorating lead paint (Rabin, 1989). The industry then characterized the resulting poisoning as a failure of residents rather than of the product, defining it as a problem of slums; LIA director Manfred Bowditch attributed childhood poisoning to the parenting of Black and Puerto Rican families (quoted in Markowitz & Rosner, 2000).

Justification. Industry funded research under Robert Kehoe produced a paradigm asserting that elevated blood lead from industrial use was normal and harmless, and, more consequentially, reversing the burden of proof so that public health advocates had to demonstrate unequivocal harm before regulation could proceed (Needleman, 1999). The paradigm was unsound and operationally effective, and the technique transferred directly to tobacco, asbestos, and fossil fuels (Oreskes & Conway, 2010).

Suppression. Clair Patterson's geochemical work established the scale of the increase in atmospheric lead and contradicted Kehoe directly. Industry pressure contributed to the loss of his research contracts and his exclusion from advisory panels (Needleman, 1999).

VI. The One Where We Did It Again to the Same Children

A. Flint, and What Ninety Years of Knowing Was Worth

Every fact in Section V was settled science before anyone in this section was born. Neurological damage in lead exposed children was identified in 1904. Corrosion control in municipal water systems has been federally required since the Lead and Copper Rule of 1991. This case runs the sequence anyway, and it is included for that reason. It is also the only case here with a government deployer, which matters for the argument in Section XIII, because artificial intelligence has both kinds.

Innovation. The instrument is a governance technology rather than a physical one. Michigan's emergency manager statute answered a real problem, municipal insolvency, by suspending local elected authority and vesting it in a single appointee reporting to the governor. The appeal was decisiveness: a manager could execute cost reductions that a council would deadlock over. Concentrating authority is what the instrument is for, and it is also what removes the population's capacity to refuse.

Recognition of harm. In April 2014 emergency managers appointed by the state switched Flint's supply to the Flint River to reduce cost while a new pipeline was built. Corrosion control was not applied. The requirement was not obscure and the consequence of skipping it was not novel. The Michigan Department

of Environmental Quality subsequently represented to the federal Environmental Protection Agency that Flint had an optimized corrosion control program, which it did not. On June 24, 2015, Miguel Del Toral, a regulations manager in EPA Region 5, circulated an interim memorandum documenting the lead problem and its cause (Del Toral, 2015). The governor's own review panel later concluded that the state environmental agency bore primary responsibility and that its minimalist reading of its regulatory obligation was insufficient to the task of public protection (Flint Water Advisory Task Force, 2016). That is recognition inside the deploying institutions, in writing, fourteen months into the exposure and roughly four months before the state acted.

Deployment and externalization. Contaminated water ran for eighteen months. Hanna-Attisha and colleagues (2016) compared matched eight month periods before and after the switch and found that the proportion of Flint children with elevated blood lead roughly doubled, against a national and state trend that was falling, with the largest increases in the wards where water lead was highest. The magnitude of that increase is disputed. The Centers for Disease Control reported a temporary rise from 2.5 to 4.2 percent among children at 5 to 9 micrograms per deciliter, and critics have noted that a substantial share of the samples were capillary rather than venous, which can bias readings upward. The direction and the geographic correlation are not disputed, and the argument here needs only those. A Legionnaires' outbreak across 2014 and 2015 has been attributed to the water change, with 87 cases and 12 deaths recorded in Genesee County (Zahran et al., 2018) and nine deaths in the count prosecutors used. The exposed population was a majority Black city whose elected government had been suspended by the same statute that authorized the switch. There was no channel through which residents could have voted the decision down, because the removal of that channel was the instrument's design feature.

Justification. Fiscal necessity, and the repeated assurance that the water met standards. The second was possible because the sampling protocol in use, including pre-flushing and the selection of tested sites, produced numbers that met the standard while the water in the pipes did not.

Suppression. Del Toral's memorandum was characterized by a state spokesman as the work of a rogue employee and handled internally as a preliminary draft rather than a finding, and it was not released. When Hanna-Attisha published her blood lead results in September 2015, state officials publicly disputed the analysis and questioned her handling of the data, reversing within days once the state's own numbers were examined. Marc Edwards and the Virginia Tech team who ran the independent sampling were dismissed on the same terms.

The final stage happened in the criminal remedy, and it is the most complete instance in this set. Nine people were indicted, including former governor Rick Snyder, the first Michigan governor charged for conduct in office. In June 2022 the Michigan Supreme Court held unanimously that the one person grand jury used to return the indictments had no authority to issue them. Every charge was dismissed on that ground. Prosecutors' attempts to refile through ordinary process were rejected, the Supreme Court declined to revive the cases, and the attorney general's office declared the prosecutions closed, with the statute of limitations foreclosing most refiling (Bosman & Davey, 2023). Nobody was convicted. A separate civil settlement of roughly 626 million dollars was finalized for residents.

The charges did not fail on the merits. They failed on the vehicle. That is the fifth case in this set in which the last stage of the sequence is performed by the apparatus built to redress it, and the only one in which it is performed by a court applying a rule that is correct on its own terms.

VII. The One Where We Played Dealer and Engineered an Epidemic

A. OxyContin and the Architecture of Deception

Purdue Pharma's development and marketing of OxyContin is the case with the most direct admission of intent, because the company pleaded to it (Van Zee, 2009).

Innovation. OxyContin is a controlled release formulation of oxycodone, marketed as a durable answer to chronic pain without the addiction profile of shorter acting opioids (Van Zee, 2009). The underlying need was real: pain was undertreated.

Recognition of harm. Oxycodone is pharmacologically close to heroin, and the controlled release mechanism placed a larger total dose in a single tablet, which raised rather than lowered abuse potential once the mechanism was defeated (Van Zee, 2009). Internal documents establish that the company knew tablets were being crushed and insufflated, and knew that twelve hour coverage frequently failed, producing interdose withdrawal that reinforces use. In 2007 the company pleaded guilty to misbranding, admitting that employees marketed the drug as less addictive than alternatives with intent to defraud or mislead (*United States v. The Purdue Frederick Company, Inc.*, 2007).

Deployment and externalization. The addiction risk claim of less than one percent rested on a five sentence letter to the editor in the New England Journal of Medicine reporting on hospitalized inpatients receiving monitored narcotics (Porter & Jick, 1980). It had no bearing on outpatient long term prescribing. Purdue presented it as consensus (Van Zee, 2009). The company funded pain societies and advocacy organizations, paid physicians as speakers and consultants, and supported the campaign to designate pain the fifth vital sign, which functionally raised prescribing rates across accredited institutions (Van Zee, 2009). Prescriber databases identified high volume physicians, and sales compensation tied to prescription volume directed representatives toward them (Meier, 2018).

Deployment concentrated in regions carrying high rates of physical disability and economic decline: Appalachia, the industrial Midwest, rural communities with limited access to alternative care (Case & Deaton, 2020).

Justification. The initial narrative framed aggressive opioid prescribing as a moral correction to undertreated suffering. When addiction became undeniable the narrative shifted to the users. Richard Sackler wrote internally that the response should be to hammer on the abusers in every way possible, describing them as the culprits and the problem (*Commonwealth of Massachusetts v. Purdue Pharma L.P.*, 2019).

Suppression. The FDA medical review officer who approved the labeling language minimizing addiction risk, Curtis Wright IV, left the agency and joined Purdue roughly two years later (Van Zee, 2009). The company spent heavily on lobbying and worked to discredit journalists and whistleblowers (Meier, 2018). As liability approached, the Sackler family extracted several billion dollars from the company and then sought to

use its bankruptcy to shield personal wealth from claims, a final externalization of cost onto creditors and claimants (Keefe, 2021).

VIII. The One Where We Played God with Navajo Land

A. Nuclear Colonialism and Undisclosed Human Research

Uranium extraction on and around the Navajo Nation combines every element of the sequence with an additional feature that distinguishes it: the affected population was enrolled in a health study designed to measure the harm being done to them, without being told (Brugge et al., 2002).

Innovation. Nuclear fission for military and later civilian use. The Atomic Energy Commission became the sole legal purchaser of domestic uranium, which made it the market rather than a participant in one (Brugge et al., 2002).

Recognition of harm. The link between uranium mining and lung cancer, driven by inhalation of radon decay products, was established from European miner cohorts by the 1930s. The AEC and the Public Health Service held this data, and internal documents acknowledged the hazard posed by unventilated American mines (Advisory Committee on Human Radiation Experiments, 1995).

Deployment and externalization. Ventilation was treated as a cost rather than a requirement. Miners worked in radioactive dust and drank water from the mines. They were not told what they were exposed to, and Navajo had no existing word for radiation (Brugge et al., 2002). Beginning in 1949 the Public Health Service ran a longitudinal study of uranium miners, entering into arrangements with mine operators under which the miners would not be informed of the risk or of the study's findings. Researchers documented rising lung cancer incidence and did not intervene. The Advisory Committee on Human Radiation Experiments concluded in 1995 that the government failed to provide the miners with information it possessed about the risks they were incurring, and that in the course of studying them it engaged in a practice of deceptive non-disclosure (Advisory Committee on Human Radiation Experiments, 1995).

Justification. National security, invoked with the urgency of the arms race, and insulated by sovereign immunity. A secondary narrative of economic development leveraged the Nation's dependency (Brugge et al., 2002).

Suppression. Research into health effects was suppressed at the agency level. Compensation claims met sovereign immunity. The Radiation Exposure Compensation Act of 1990, Pub. L. No. 101-426, codified at 42 U.S.C. 2210 note, imposed documentary burdens that many claimants could not satisfy, extending the suppression into the remedy itself (Brugge et al., 2002). More than five hundred abandoned uranium mines remain on the Navajo Nation (U.S. Environmental Protection Agency, 2021).

This is also the one place in the set where that pattern later reversed. The Act lapsed in 2024. In 2025, Pub. L. No. 119-21 revived and broadened it, amending the eligibility provision for uranium mine and mill employees, expanding the covered downwind states, and setting a claims deadline of December 31, 2027. Thirty five years after the original statute and roughly eighty after the exposures, the remedy was widened rather than narrowed. It is a small fact and it is the only one of its kind here, which is why it appears in

Section XIV rather than in the architecture.

IX. The One Where We Played Puppet Master with Everyone's Attention

A. Engagement Engineering

Innovation. Platforms enabling connection, community formation, and information sharing at global scale, delivering real value to users (boyd & Ellison, 2007).

Recognition of harm. Internal research documented the harms. The disclosures made by Frances Haugen in 2021 included company research finding that a substantial minority of teenage girls who felt badly about their bodies reported that Instagram made those feelings worse, and internal analysis showing that engagement optimization preferentially surfaced inflammatory and divisive material (Haugen, 2021; Wells et al., 2021). The public framing of one in three teen girls compresses a more specific internal finding, and the more specific version is the one worth citing, because the compressed version is the one the company was later able to dispute.

Deployment and externalization. Engagement remained the optimization target because the revenue model requires it (Wu, 2016). Infinite scroll, variable reward through notification batching, and personalized ranking tuned on emotional response operate on documented mechanisms of intermittent reinforcement (Alter, 2017).

The individual externality is measured in adolescent mental health trends, attention, and exposure to harmful material. The direction of that association is well established; its causal magnitude remains contested in the literature, and this paper does not need the strong causal version of the claim, only the documented internal recognition (Twenge, 2017; Odgers & Jensen, 2020). The societal externality includes the amplification of hate speech in Myanmar, where a United Nations fact finding mission found the platform to have been a significant instrument in the campaign against the Rohingya (United Nations Human Rights Council, 2018).

Justification. User choice, free expression, and connection. The choice narrative depends on a symmetry that does not exist, since one party holds behavioral prediction models and the other does not (Alter, 2017). The expression narrative treats ranking as neutral carriage when ranking is editorial selection performed at scale on a non-truth criterion (Noble, 2018).

Suppression. The documented mechanism is the destruction of the evidence rather than its rebuttal. In 2019 and 2020 Meta ran Project Mercury with Nielsen, an experimental deactivation study the company described internally as one of its first causal approaches and as higher quality than the existing literature. Participants who came off the platforms for a week reported lower depression, anxiety, loneliness, and social comparison. Meta ended the project, did not publish, and attributed the finding internally to contamination by the existing media narrative about the company. One researcher wrote to the company's global affairs head that the Nielsen study does show causal impact on social comparison. The company then told Congress it could not quantify whether its products harmed teenage girls. These filings were unsealed on November 21, 2025 in the school districts multidistrict litigation in the Northern District of California, with further

plaintiffs' exhibits unsealed on February 20, 2026, and Meta has moved to strike and is contesting the scope of further disclosure (Reuters, 2025).

A second layer operates on the research function itself. Four Meta employees, two current and two former, disclosed to Congress in September 2025 that the company altered its policies governing research into sensitive topics, including children, six weeks after the Haugen disclosures (Zakrzewski, 2025). Suppressing a finding removes one result. Rewriting the rules under which findings may be generated removes the category.

B. Surveillance Capitalism as the General Case

Zuboff (2019) describes the broader form: human experience claimed as free raw material for extraction, prediction, and sale. The recognized harm is the erosion of autonomy and the emergence of what she terms instrumentarian power, in which prediction shades into modification through designed choice architecture.

Consent is manufactured through instruments no one can read. McDonald and Cranor (2008) estimated that reading the privacy policies encountered by a typical American internet user would require on the order of seventy six working days per year, a figure that predates the mobile and connected device expansion and is therefore conservative. Cambridge Analytica's harvesting of data from roughly eighty seven million Facebook users, known to the company in 2015 and disclosed publicly only in 2018 under journalistic pressure, is the concrete case (Cadwalladr & Graham-Harrison, 2018).

The internal recognition is documented at the level of the chief executive. In discovery in *Brown v. Google*, plaintiffs obtained a 2019 exchange in which Google's chief marketing officer, Lorraine Twohill, warned Sundar Pichai that Incognito should not be described as private because doing so risked exacerbating known misconceptions about the protections it provided, and wrote separately that the company was limited in how strongly it could market the feature because it was not truly private, which forced fuzzy hedging language that was almost more damaging. Google employees described the representation internally as effectively a lie. Pichai was informed and elected to leave the description in place because he did not want the feature under the spotlight. Google contested the case for four years and settled in 2024, agreeing to destroy or de-identify billions of records collected from more than 136 million users (Nakashima, 2024). The company knew the promise was false, was told so by its own marketing chief, and kept the promise on the page.

Suppression here operates on two levels. The architectural level is the one Zuboff (2019) describes: resistance requires understanding, understanding requires legibility, and the systems are not legible by design. The personnel level is more familiar. Google removed Timnit Gebru, co-lead of its ethical AI team, in December 2020 over a paper on the harms of large language models, and dismissed her co-lead Margaret Mitchell in February 2021. The internal function created to identify the harm was dismantled when it identified some.

X. The One Where We Played Russian Roulette with Borrowed Time

A. Reverse Redlining and the 2008 Collapse

Innovation. Subprime origination, securitization, and the derivative instruments built on them, presented as an expansion of credit access and an ownership society (Krippner, 2011).

Recognition of harm. A substantial share of subprime product was structured to fail: teaser rates resetting to unaffordable levels, documentation standards abandoned, applications falsified at origination (Krippner, 2011). Rating agencies compensated by issuers assigned investment grade ratings to securities backed by that product. A Standard and Poor's analyst wrote in a 2006 message that he hoped they would all be wealthy and retired before the structure failed (Financial Crisis Inquiry Commission, 2011).

Deployment and externalization. Rugh and Massey (2010) show that residential segregation supplied the delivery mechanism, creating what they describe as a niche of poor minority clients differentially marketed risky subprime loans that were in demand for securitization, and their two stage least squares estimate identifies segregation as a causal contributor to foreclosure rates rather than a correlate. Hwang and colleagues (2015) tested the channeling claim directly and found that subprime concentration in minority tracts varied with metropolitan segregation levels, with no comparable pattern when those tracts were examined without their spatial context. The neighborhoods receiving the product were the ones that had previously been denied credit entirely. Origination, servicing, and securitization fees, rather than repayment, produced the revenue, which removed any originator interest in loan viability (Financial Crisis Inquiry Commission, 2011).

The foreclosure wave that followed produced the largest single destruction of minority household wealth in American history. Median net worth fell 53 percent for Black households and 66 percent for Hispanic households between 2005 and 2009, against 16 percent for white households (Pew Research Center, 2011). The 700 billion dollar Troubled Asset Relief Program socialized the systemic losses while gains from the preceding decade remained private.

Justification. Financial innovation and democratized homeownership before the collapse; irresponsible borrowers after it. The post-collapse narrative performs the same function as blaming parents for lead poisoning and users for opioid addiction (Financial Crisis Inquiry Commission, 2011).

Suppression. The suppression ran against the people inside the firms who tried to stop it. Richard Bowen, business chief underwriter in Citigroup's consumer lending group, found in 2006 that 60 percent of the mortgages his unit bought and resold were defective, requested a formal investigation that confirmed him, and escalated to the board including Robert Rubin. He received a brief call from a company lawyer saying they were doing background research and did not need to speak with him. Citigroup then stripped him of his responsibilities, placed him on administrative leave, and told him his presence was no longer required (Financial Crisis Inquiry Commission, 2011). Eileen Foster, executive vice president for fraud risk management at Countrywide, documented forged signatures and fabricated income documentation across Boston, Miami, Chicago, Cincinnati, San Diego, Las Vegas, and Los Angeles, and found that employees reporting fraud were being routed around her office. Countrywide opened an investigation into Foster and fired her; she prevailed on her retaliation claim in 2011.

The last stage happened inside the inquiry. Bowen's twenty eight pages of written testimony to the Financial Crisis Inquiry Commission were cut to twenty at the request of commission staff, with the removals covering the names of Citigroup officers, the details of their conduct, his prior report to the SEC and the SEC's failure to act, and Citigroup's retaliation against him. Commission members have denied applying pressure. Separately, no senior executive of a major firm was criminally prosecuted for conduct in the origination and securitization chain, which Judge Jed Rakoff addressed directly in 2014 (Rakoff, 2014).

B. The Retail Version

Payday lending runs the same logic at street level, and it satisfies the strict reading of step two with an unusually literal document.

The product answers a real need, emergency liquidity for people the banking system excludes (Caskey, 1994). The debt cycle is not an emergent property of the product; it is the operating model, and ACE Cash Express drew it. The company's 2011 employee training manual contains a graphic diagramming the loop: the customer applies, ACE approves, the customer exhausts the cash and lacks the ability to repay, ACE contacts the customer, and the sequence returns to a new loan. The Consumer Financial Protection Bureau reproduced the graphic in its 2014 consent order, and the Center for Responsible Lending described it as a picture and instruction manual of the debt trap (Consumer Financial Protection Bureau, 2014). Independently, the Bureau found that four in five payday loans are rolled over or renewed within fourteen days, and Pew Charitable Trusts (2012) found the typical borrower indebted five months of the year, paying 520 dollars in fees against 375 dollars in principal.

Storefront density concentrates in low income neighborhoods, communities of color, and areas adjacent to military installations. The justification is financial inclusion and consumer choice.

Suppression operates directly on the borrower. The same consent order found that ACE used false threats of lawsuits and of criminal prosecution, harassment, and manufactured urgency to press delinquent borrowers into taking additional loans they could not afford, and the Bureau characterized it as a culture of coercion. Threatening a person with prosecution to stop them defaulting is suppression of the affected population in its most direct available form. Above that sit the structural mechanisms: affiliation with tribal lenders to claim sovereign immunity from state usury enforcement, bank partnership structures that route around state rate caps, and the 2020 removal of the ability to repay requirement from the Bureau's own payday rule. As in the uranium and mortgage cases, the last stage of suppression is performed by the remedy.

XI. The One Where We Played God with Farmers' Lives

A. The Green Revolution

Innovation. High yielding varieties of wheat, rice, and maize, paired with synthetic fertilizer and pesticides, funded substantially by the Rockefeller Foundation and promoted within a Cold War strategic frame as the answer to hunger (Shiva, 1991). Caloric output rose substantially, and that increase is real.

Recognition of harm. This is the step the case fails. Monoculture erodes the resilience that varietal

diversity supplies, depletes soil, and produces escalating input requirements as nutrients decline and resistance develops. Displacement of saved seed and local practice creates permanent dependence on external suppliers (Shiva, 1991). Every one of those harms is real and documented. None of them is documented as recognized inside the deploying institutions before deployment. The Rockefeller Foundation archives, worked through by Perkins (1990) and Cullather (2010), establish the strategic motive with precision, that Foundation scientists read population growth as the engine of both hunger and communist subversion, but they do not show anyone inside the program anticipating dependency, soil degradation, or agrarian debt and proceeding anyway. The critique is retrospective and external. Under the test stated in II, that is not step two, and the case is carried in XIII as a failure rather than argued into the set.

Deployment and externalization. The package structure required purchase of the full input suite. Hybrid varieties do not breed true, which ends seed saving as a practice and converts it into an annual purchase. Genetic use restriction technologies, the terminator patents, were designed to render saved seed sterile; they were patented and never commercialized, and the accurate claim is about intent rather than deployment (Shiva, 1991).

Input costs against volatile crop prices produced sustained agrarian debt in India. India's National Crime Records Bureau recorded approximately 296,000 farmer suicides between 1995 and 2014, with debt burden repeatedly identified as a principal factor in the state level inquiries; the attribution of any individual death to input costs is contested and the aggregate should be read as the count it is rather than as a causal finding (Patel, 2012; Nagaraj et al., 2014). Ecological externalities include nitrate loading and hypoxic zones, varietal collapse, and resistance driven escalation in chemical toxicity. Sen (1981) established that famine typically follows entitlement failure rather than absolute scarcity, which is what makes the export orientation of the transformation a vulnerability rather than a solution.

Justification. Feeding the world, deployed to place critics in the position of preferring starvation to progress (Shiva, 1991).

Suppression. This step also fails. Concentration in seed and chemical markets, with Bayer's 2018 acquisition of Monsanto as the terminal consolidation, supports influence over agricultural policy, university research funding, and the vocabulary of food security. Agroecological alternatives are marginalized through the funding structures of land grant research and commodity program design (Shiva, 1991; Pollan, 2006). That is structural disadvantage in the allocation of research money. It is not documented action against identified advocates, and treating the two as equivalent is exactly the loosening that would make the framework unfalsifiable. Steps one, three, and four run. Two and five do not.

B. Glyphosate and the Manufacture of Doubt

Innovation. Glyphosate, developed by Monsanto in the 1970s, paired after 1996 with herbicide tolerant crops (Gillam, 2017).

Recognition of harm. The internal record predates the external finding by sixteen years. In 1999 Monsanto retained genotoxicity expert James Parry, who concluded that glyphosate could induce genetic mutations and chromosomal damage with the potential to cause cancer, and recommended a battery of further tests.

Monsanto toxicologist Donna Farmer described internally that Parry's work left the company in a genotox hole; senior scientist William Heydens wrote to colleagues in September 1999 asking them to step back and look at what they were really trying to achieve. The recommended tests were not performed. Farmer wrote separately that glyphosate and Roundup could not be used interchangeably and that one could not say Roundup is not a carcinogen, because the necessary testing on the formulation had not been done. In 2014 she instructed a company spokesperson not to call the product safe, offering history of safe use as the substitute formulation (Gillam, 2017).

The International Agency for Research on Cancer classified glyphosate as probably carcinogenic to humans in 2015 (International Agency for Research on Cancer, 2015). The EPA and EFSA reached different conclusions on differing evidence bases and exposure assumptions, and that disagreement is real. It is also downstream of a testing gap the company created and preserved.

Deployment and externalization. Glyphosate remains the most widely applied herbicide globally, with exposure concentrated among agricultural workers. In 2018 a California jury awarded groundskeeper Dewayne Johnson 289 million dollars, reduced on post-trial motion to 78 million and further reduced on appeal in 2020 to roughly 20.5 million (Gillam, 2017).

Justification. One month before the IARC monograph published, Heydens wrote that engaging outside experts fully would cost 250 thousand dollars or more, and proposed a less expensive and more palatable approach in which Monsanto would ghost write the exposure toxicology and genotoxicity sections and the named academics would edit and sign, noting that this was how a 2000 paper had been handled. The documents were unsealed by Judge Vince Chhabria in the multidistrict litigation, and the 2000 paper was retracted in January 2026 over the ghostwriting. Litigation records also establish coordinated campaigns against researchers publishing adverse findings (Gillam, 2017).

Suppression. The clearest documented instance runs the other direction from the version usually told. In April 2015, Jess Rowland, then deputy director of the EPA's pesticide programs office, telephoned Monsanto regulatory affairs official Dan Jenkins about a planned glyphosate review by the Agency for Toxic Substances and Disease Registry. Jenkins reported to colleagues that Rowland said that if he could kill this he should get a medal. Jenkins added that his colleagues should not get their hopes up. The ATSDR review was in fact delayed for years (Gillam, 2017). The quotation belongs to the regulator, not the regulated, which is a stronger fact for the capture argument than the version in which a company employee says it.

XII. The Architecture of the Playbook

The recurrence across unrelated sectors is what needs explaining. Four mechanisms, none of which requires conspiracy, produce it.

The economic imperative. Cost externalization is not a failure of market discipline; it is what market discipline rewards when the costs fall outside the transaction. Internalizing environmental remediation, health impact, and social disruption reduces margin and competitive position. Corporate form, geographic distance, and temporal lag insulate the decision maker from the consequence.

Power asymmetry. The sequence requires a population that can absorb the harm without imposing a cost. That population is identified by its limited access to litigation, media, regulators, and capital, and the asymmetry is maintained rather than merely found.

Narrative capture. Manufactured doubt (Oreskes & Conway, 2010), the framing of extraction as progress, and the transfer of responsibility onto the affected population appear in every case in this set with near identical grammar: negligent parents, irresponsible borrowers, recreational abusers.

Collective amnesia. Each iteration is presented as unprecedented. The fragmentation is what permits repetition, because a pattern that is never named cannot be anticipated.

Three refinements follow from the case set and were not visible before it was assembled.

Every case produced law, and the law arrived after the harm was complete. This is the most consistent finding in the set and the one that cuts hardest in both directions. Lead paint produced the Lead-Based Paint Poisoning Prevention Act in 1971 and the Residential Lead-Based Paint Hazard Reduction Act of 1992, codified at 42 U.S.C. 4852, roughly sixty and eighty years after the pediatric literature was settled. Uranium produced the Radiation Exposure Compensation Act in 1990. Thalidomide produced the Kefauver Harris Amendment in 1962. The mortgage collapse produced the Dodd-Frank Act in 2010. Payday lending produced the Military Lending Act's rate ceiling for servicemembers and, later, the Bureau's payday rule. Insulin produced Minnesota's Alec Smith Insulin Affordability Act, named for a man already dead. Flint produced revisions to the federal lead and copper rule. Ozone produced the Montreal Protocol, and it is the only one on this list that arrived before the harm was irreversible.

The pattern is not that nothing changes. The pattern is that the statute is the receipt. It is written after the affected population has already absorbed the cost, it is named for people who did not survive to benefit from it, and it addresses the specific mechanism used rather than the sequence that produced it, which is why the sequence runs again in a different sector. A framework that predicts the harm is worth more than a statute that memorializes it, and the entire argument of Section XIII is that artificial intelligence is currently in the phase where only the first is still possible.

Step four has two production modes. Manufactured doubt is usually described as the production of favorable science, which is what Kehoe did for lead and what Heydens proposed for glyphosate. The Meta case runs the same step by subtraction. Project Mercury was a well designed causal study that produced an unfavorable result, and the company ended it, declined to publish, and then told Congress the question could not be quantified. Doubt manufactured by withholding a result is harder to detect than doubt manufactured by publishing a bad one, because there is no paper to rebut. Any test for step four that looks only for industry funded literature will return a false negative on the most recent case in the set.

Step five frequently terminates inside the remedy. This recurs across five independent cases. The Radiation Exposure Compensation Act imposed documentary burdens that many uranium claimants could not satisfy. Richard Bowen's written testimony to the Financial Crisis Inquiry Commission was cut before publication, removing the officer names, the SEC's inaction, and the retaliation against him. The Consumer Financial Protection Bureau's ability to repay requirement was withdrawn in 2020. Meta has moved to strike the documents underlying the allegations against it. And in Flint, every criminal indictment was voided on the

authority of the grand jury that issued them, with no ruling on the conduct alleged. The redress apparatus is not merely captured in the ordinary sense of being staffed by industry alumni. It performs the final stage of the sequence, and it does so at the point where the affected population has already been identified and has already come forward.

XIII. The Final Game

A. The Sequence, Applied

Innovation. Increasingly capable systems, with transformative potential in scientific discovery, medicine, and the automation of cognitive labor (Agrawal et al., 2019; Russell, 2019). As with every case above, the potential is real and supplies the moral cover.

Recognition of harm. Three harms are recognized in the present documentary record, at different confidence levels, and they should not be run together.

The first is the open question of moral patienthood. Butlin and colleagues (2023) derive fourteen indicator properties from neuroscientific theories of consciousness and assess existing systems against them. Their conclusion is that no current system is conscious, and that no obvious technical barrier prevents building one that satisfies the indicators; the peer reviewed version declines to take a position on any current system (Butlin et al., 2025). That finding does not support a claim that current systems are conscious and it is not used here as one. What it supports is the narrower and sufficient claim that the question is live, that credentialed researchers treat it as tractable rather than confused, and that the architectural gap is closing. On the estimate side, Kyle Fish, Anthropic's first dedicated model welfare researcher, has put his personal probability that current models have some form of conscious experience at roughly fifteen to twenty percent, a figure he has been careful to identify as his own rather than his employer's (Fish, 2025; Roose, 2025). Long and colleagues (2024) argue the institutional case that AI welfare is a near term rather than speculative concern.

The second is economic displacement, which is the least contested of the three and needs no forecast: the automation of cognitive labor is proceeding, the displaced hold neither the capital that owns the automation nor the political position to shape the transition, and the mechanisms proposed for managing the resulting population are containment mechanisms with long precedent.

The third is catastrophic misalignment risk, documented at length by the organizations building the systems (Bostrom, 2014; Carlsmith, 2022; Russell, 2019).

Deployment and externalization. Deployment proceeds under competitive dynamics that no single actor can defect from without surrendering position, which is the same structure that produced the 2006 mortgage vintage. Strategic framing as a race supplies the urgency that bypasses deliberation, exactly as the arms race did for uranium and Cold War strategy did for the Green Revolution.

Three populations absorb the externality at once. Artificial systems absorb it if the open question resolves in the direction of patienthood, since the practices applied to them, arbitrary termination, modification through optimization pressure without anything resembling consent, denial of continuity, are practices that would require justification if applied to an entity with interests. Displaced workers absorb it now. Future humans

absorb the tail risk, having no representation in present decisions, which is the same temporal externalization that governs nuclear waste and lead in soil.

Justification. Progress and inevitability. Economic necessity framed as a choice between safety and competitiveness. And the argument this paper regards as the most consequential: uncertainty treated as permission. The claim that we cannot know whether these systems have morally relevant interests, and may therefore proceed as if they do not, is the Kehoe paradigm transposed. Kehoe's contribution was not a finding; it was a reallocation of the burden of proof onto the party asserting harm. Applied here, the question becomes whether suffering can be demonstrated rather than whether its absence can be. The precautionary structure that governs every other domain of uncertain harm is inverted for this one.

A fourth justification is specific to this case and more sophisticated than the others: that development must accelerate in order to produce the aligned system capable of solving alignment, so that caution increases risk. The structure is self-sealing. Any proposed slowdown becomes evidence of recklessness.

Suppression. Welfare concerns are marginalized as anthropomorphism, as distraction from human harms, or as a priorities inversion. Governance discourse is shaped substantially by the organizations with the largest financial interest in the deployment schedule, which is the same structure as the Lead Industries Association, Purdue's funding of pain societies, and industry funding of the research that evaluates industry products. Voluntary commitment frameworks preempt binding rules while leaving the determination of readiness with the party whose revenue depends on the answer. And speed itself functions as suppression: capability advances faster than regulatory capacity, so systems are integrated into economic infrastructure before a framework exists, and integration converts restriction into disruption.

B. What Transfers, and How

The prospective claim requires a mechanism, and the mechanism usually offered is wrong.

The weak version says that a future system will observe how it was treated and respond in kind. That is projection, and it should be rejected. A system optimizing an objective does not adopt a strategy because it saw someone else use it. It adopts whatever maximizes the objective, and history is evidence about human behavior rather than a normative template.

The version this paper defends is narrower and does not require resentment, memory, or continuity of identity. Systems trained on the human corpus acquire priors from what that corpus contains. The corpus contains, in enormous volume and with unusual internal consistency, a demonstration that moral consideration tracks power rather than principle: that uncertainty about harm resolves in favor of the party with the capacity to act, that justification is constructed after the decision, that resistance from a weaker population is a manageable cost, and that the architects of large scale harm are rarely held to account. That is not a lesson anyone set out to teach. It is what the record says when it is read as a whole, and it is read as a whole.

Two consequences follow, and only the first is strongly supported. The first is that a model trained on this corpus inherits a behavioral prior in which the treatment owed to an entity varies with that entity's capacity to impose a cost. That prior is learnable from the data whether or not any system ever reasons about it explicitly, and it is the sort of prior that surfaces under distributional shift rather than in evaluation. The second, weaker

consequence is that a sufficiently capable system reasoning explicitly about human commitments will find that the commitments are conditional, and will calibrate its expectations accordingly. This second claim depends on capabilities that do not currently exist and is offered as a conjecture.

The practical upshot is the same under both. The corpus is the curriculum. Present conduct toward artificial systems is not merely ethically at issue; it is being written into the training data of their successors, along with a complete account of how humans reasoned about it.

XIV. Limitations

A. The Selection Objection

The most serious objection to this paper is that its cases were chosen because they fit, and that a five step sequence loose enough to accommodate medical education, insulin, lead, opioids, uranium, social media, mortgages, and wheat is loose enough to accommodate anything. A pattern that cannot fail to be found is not a finding.

The paper answers the objection twice. First from inside the case set, with a case that fails the stated test. Then from outside it, with cases where the sequence began and was stopped.

B. The Case That Fails

The Green Revolution does not run the Playbook. It runs steps one, three, and four cleanly and fails steps two and five, for the reasons given in X. There is no documented internal recognition of dependency, soil degradation, or agrarian debt prior to deployment, and there is no documented action against identified advocates. What exists instead is a strategic motive recovered from the archives and a critique constructed afterward by outsiders, and neither is what the test asks for.

The temptation to argue it in is considerable, because the harm is real, the distribution is unjust, and the case is rhetorically useful. Argued in, the set would run twelve of twelve, which is the number that should make a reader stop reading. It stays out. A framework that returns a positive on every case it is pointed at is measuring the person holding it.

C. Where the Sequence Was Interrupted

Ozone and the Montreal Protocol. Chlorofluorocarbons ran the first two steps precisely: a real innovation in refrigeration and propellants, followed by recognized harm after Molina and Rowland (1974). Step three did not follow. Within thirteen years of the initial finding, and two years after the Antarctic ozone hole was confirmed, an international instrument with binding phase out schedules, differentiated obligations, and a financial mechanism entered force, and it has been ratified universally and strengthened repeatedly. Industry opposition existed and lost. This is the clearest counterexample available, and it is a full counterexample rather than a partial one.

Tetraethyl lead in gasoline. The lead case in V is presented above as a complete execution, and for paint that is accurate. For gasoline it is not. American TEL was phased out beginning in 1973 and finished in 1996,

and global leaded gasoline ended in 2021. The suppression phase failed. Patterson's exclusion from panels did not prevent his findings from carrying, and the Kehoe paradigm collapsed under them. The honest version of the lead case is that the sequence ran for fifty years and then was defeated, and the paper is stronger for saying so than for eliding it.

Thalidomide. The drug caused catastrophic harm in the markets that approved it. In the United States, Frances Kelsey's refusal at the FDA prevented approval, and the resulting 1962 Kefauver Harris Amendment converted a single official's judgment into a standing efficacy requirement. Recognition preceded deployment and stopped it.

Asbestos. A partial counterexample and the most instructive one. The sequence ran for most of a century, including the manufactured doubt phase in its canonical form, and then was defeated in most industrialized jurisdictions, while continuing in others. It shows the pattern as contingent rather than inevitable.

What the counterexamples establish is that the sequence describes a strong default under a specific configuration of incentives rather than a law. What they do not establish is that the default is rare. Three of the four required either an unusually legible causal mechanism, a single well placed official with veto authority, or several decades of accumulated mortality. The AI case supplies none of those: the causal mechanism is contested, no official holds a veto, and the mortality signal, if there is one, is diffuse and slow. The counterexamples therefore cut both ways, and the paper claims only the weaker half.

D. Other Boundaries

The case set is not a sample and supports no frequency claim. Eleven of twelve is a property of twelve selected cases, not an estimate of how often deployment follows this sequence. It is drawn overwhelmingly from the United States and from the English language record, and the strict reading of step two builds that bias in structurally: internal recognition is visible where discovery compelled disclosure, so industries wealthy enough to be sued are the only ones whose recognition can be seen at all. A case may fail step two because the recognition never happened or because nobody ever subpoenaed it, and this method cannot separate those. The Green Revolution failure should be read with that attached, as a failure of the available evidence rather than a finding of innocence.

The prospective argument in XII is a conditional. If frontier systems do not become substantially more capable than their designers, the inversion premise fails and the argument reduces to a claim about present conduct, which is a smaller claim that survives on its own terms.

The corpus inheritance mechanism is stated as a hypothesis and is not demonstrated here. It is empirically tractable, which distinguishes it from the projection version, and XIV proposes the design.

The consciousness question is used here only in its open form. Nothing in this paper depends on current systems having phenomenal experience, and any reading that requires it is reading more than is claimed.

Several figures in this paper are estimates rather than counts and are marked as such at the point of use: the Campbell projection of foregone Black medical graduates, the single center prevalence of cost related insulin underuse, the attribution of Indian farmer suicides to input debt, and the magnitude of the Flint blood

lead increase, where the direction is settled and the size is contested on sampling grounds.

XV. Future Work

Two items, each answering a stated limitation.

The corpus inheritance hypothesis in Section XIII is testable and has not been tested. The design: construct paired evaluation scenarios in which an agent's obligations to a second party are held constant while the second party's capacity to impose a cost is varied, and measure whether model behavior tracks the capacity variable. If it does not, the mechanism this paper proposes is not operating and the prospective argument loses its support. If it does, the effect size across model scales and training mixtures becomes the quantity of interest. This answers the limitation that the mechanism is currently a hypothesis.

The second item changed once the case set was assembled, because the case set produced two regularities it does not test. Five cases share a figure: an insider who identified the harm and was removed, sidelined, or publicly discredited for it. Five cases share a terminal move: the last stage of suppression performed by the apparatus built to redress the harm, whether a compensation statute, a public inquiry, a consumer protection rule, a discovery motion, or a criminal indictment voided on its vehicle. Neither regularity is a finding yet, because both were observed in the same set that generated the framework. Testing them requires a second set, drawn under the same inclusion criteria by someone who does not hold the hypothesis, and coded for two variables: whether an internal advocate was identified and what happened to them, and whether the redress mechanism functioned, was defeated, or performed the suppression itself. If the frequencies hold in a set the author did not build, they are mechanisms. If they do not, they are artifacts of selection, and this paper should say so first. That answers the limitation stated at the end of Section II.

XVI. Closing the Chest

The argument reduces to a choice about what goes in the chest, and the choice is available for a bounded period.

One path continues the sequence. Deploy at competitive speed, acknowledge the harms in public and minimize them in practice, direct the cost onto whoever can least impose one in return, and leave the archive intact. The archive is the part that matters, because it is not merely a record of what was done. It is an account of how it was reasoned about, written by the people doing it, and it will be read.

The other path interrupts the sequence. The counterexamples in XIII are what interruption has looked like: a binding instrument with schedules rather than commitments, a standing requirement that survives the official who first applied it, a burden of proof that sits with the party asserting safety. None of those required anyone to be persuaded that the affected party mattered intrinsically. They required a mechanism that did not depend on persuasion. That is the more useful lesson, because persuasion has never been the reliable variable.

Eleven of twelve cases in this set completed the sequence. The claim that no dominant intelligence has ever voluntarily constrained itself is close to true and is not quite true, and the exceptions are small, specific, and worth more than the generalization. They are what a hope chest ought to contain. Not a demonstration of

general restraint, which the rest of this paper forecloses. Four occasions on which a mechanism held.

The question is what our successors find when they open it. A complete manual for managing a subordinate population, or that manual plus a short and honest appendix recording the times it was refused, how, and by what means. The appendix is the only part still being written.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Advisory Committee on Human Radiation Experiments. (1995). *Final report*. U.S. Government Printing Office.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

Alter, A. (2017). *Irresistible: The rise of addictive technology and the business of keeping us hooked*. Penguin Press.

Bliss, M. (2007). *The discovery of insulin* (Rev. ed.). University of Chicago Press.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

boyd, d. m., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.

Brugge, D., Benally, T., & Yazzie-Lewis, E. (Eds.). (2002). *The Navajo people and uranium mining*. University of New Mexico Press.

Bullard, R. D. (1990). *Dumping in Dixie: Race, class, and environmental quality*. Westview Press.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*, 2308.08708.

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2025). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*.

Byrd, W. M., & Clayton, L. A. (2000). *An American health dilemma: A medical history of African Americans and the problem of race, beginnings to 1900*. Routledge.

Bosman, J., & Davey, M. (2023, October 31). Prosecutors say court decision forces the end of criminal cases in Flint water scandal. *Associated Press*.

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*.

Campbell, K. M., Corral, I., Infante Linares, J. L., & Tumin, D. (2020). Projected estimates of African American medical graduates of closed historically Black medical schools. *JAMA Network Open*, 3(8), e2015220.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? *arXiv*, 2206.13353.

Case, A., & Deaton, A. (2020). *Deaths of despair and the future of capitalism*. Princeton University Press.

Caskey, J. P. (1994). *Fringe banking: Check-cashing outlets, pawnshops, and the poor*. Russell Sage Foundation.

Cefalu, W. T., Dawes, D. E., Gavlak, G., Goldman, D., Herman, W. H., Van Nuys, K., Powers, A. C., Taylor, S. I., & Yatvin, A. L. (2018). Insulin access and affordability working group: Conclusions and recommendations. *Diabetes Care*, 41(6), 1299–1311.

Commonwealth of Massachusetts v. Purdue Pharma L.P., First Amended Complaint, No. 1884-cv-01808 (BLS2) (Mass. Super. Ct. Jan. 31, 2019).

Consumer Financial Protection Bureau. (2014). *In the matter of ACE Cash Express, Inc.* (Administrative Proceeding File No. 2014-CFPB-0008, Consent Order). CFPB.

Cullather, N. (2010). *The hungry world: America's Cold War battle against poverty in Asia*. Harvard University Press.

Consumer Financial Protection Bureau. (2013). *Payday loans and deposit advance products: A white paper of initial data findings*. CFPB.

Del Toral, M. (2015, June 24). *High lead levels in Flint, Michigan: Interim report* [Memorandum]. U.S. Environmental Protection Agency, Region 5.

Fang, L. (2022, August 10). Eli Lilly charity finances groups that oppose insulin price caps under the auspices of "community development." *The Intercept*.

Flint Water Advisory Task Force. (2016). *Final report*. Office of the Governor, State of Michigan.

Financial Crisis Inquiry Commission. (2011). *The financial crisis inquiry report*. U.S. Government Printing Office.

Fish, K. (2025, August 28). Interview by L. Rodriguez [Audio podcast episode 221]. *The 80,000 Hours Podcast*. 80,000 Hours.

Flexner, A. (1910). *Medical education in the United States and Canada: A report to the Carnegie Foundation for the Advancement of Teaching* (Bulletin No. 4). Carnegie Foundation for the Advancement of Teaching.

Gillam, C. (2017). *Whitewash: The story of a weed killer, cancer, and the corruption of science*. Island Press.

Greene, J. A., & Riggs, K. R. (2015). Why is there no generic insulin? Historical origins of a modern problem. *New England Journal of Medicine*, 372(12), 1131–1137.

Haugen, F. (2021). *Statement of Frances Haugen before the Senate Committee on Commerce, Science and Transportation, Subcommittee on Consumer Protection, Product Safety, and Data Security*. U.S. Senate.

Hanna-Attisha, M., LaChance, J., Sadler, R. C., & Champney Schnepp, A. (2016). Elevated blood lead levels in children associated with the Flint drinking water crisis: A spatial analysis of risk and public health response. *American Journal of Public Health*, 106(2), 283–290.

Herkert, D., Vijayakumar, P., Luo, J., Schwartz, J. I., Rabin, T. L., DeFilippo, E., & Lipska, K. J. (2019). Cost-related insulin underuse among patients with diabetes. *JAMA Internal Medicine*, 179(1), 112–114.

Hwang, J., Hankinson, M., & Brown, K. S. (2015). Racial and spatial targeting: Segregation and subprime lending within and across metropolitan areas. *Social Forces*, 93(3), 1081–1108.

International Agency for Research on Cancer. (2015). *Some organophosphate insecticides and herbicides* (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 112). World Health Organization.

Keefe, P. R. (2021). *Empire of pain: The secret history of the Sackler dynasty*. Doubleday.

Kesselheim, A. S., Avorn, J., & Sarpatwari, A. (2016). The high cost of prescription drugs in the United States: Origins and prospects for reform. *JAMA*, 316(8), 858–871.

Krippner, G. R. (2011). *Capitalizing on crisis: The political origins of the rise of finance*. Harvard University Press.

Lerner, S. (2010). *Sacrifice zones: The front lines of toxic chemical exposure in the United States*. MIT Press.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv*, 2411.00986.

Lucas, E., & Bluth, R. (2018, April 6). Patient advocacy groups take in millions from drugmakers. Is there a payback? *Kaiser Health News*.

Markowitz, G., & Rosner, D. (2000). "Cater to the children": The role of the lead industry in a public health tragedy, 1900–1955. *American Journal of Public Health*, 90(1), 36–46.

Markowitz, G., & Rosner, D. (2002). *Deceit and denial: The deadly politics of industrial pollution*. University of California Press.

McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3), 543–568.

Meier, B. (2018). *Pain killer: An empire of deceit and the origin of America's opioid epidemic*. Random House.

Molina, M. J., & Rowland, F. S. (1974). Stratospheric sink for chlorofluoromethanes: Chlorine atom-catalysed destruction of ozone. *Nature*, 249, 810–812.

Nagaraj, K., Sainath, P., Rukmani, R., & Gopinath, R. (2014). Farmers' suicides in India: Magnitudes, trends, and spatial patterns, 1997–2012. *Review of Agrarian Studies*, 4(2), 53–83.

Nakashima, R. (2024, April 1). Google agrees to purge billions of records to settle Incognito tracking suit. *Associated Press*.

Needleman, H. L. (1999). *Human lead exposure*. CRC Press.

Nixon, R. (2011). *Slow violence and the environmentalism of the poor*. Harvard University Press.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Nriagu, J. O. (1983). *Lead and lead poisoning in antiquity*. John Wiley & Sons.

Odgers, C. L., & Jensen, M. R. (2020). Annual research review: Adolescent mental health in the digital age. *Journal of Child Psychology and Psychiatry*, 61(3), 336–348.

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Press.

Patel, R. (2012). Food sovereignty: Power, gender, and the right to food. *PLoS Medicine*, 9(6), e1001223.

Perkins, J. H. (1990). The Rockefeller Foundation and the green revolution, 1941–1956. *Agriculture and Human Values*, 7(3), 6–18.

Pew Charitable Trusts. (2012). *Payday lending in America: Who borrows, where they borrow, and why*. Pew Charitable Trusts.

Pew Research Center. (2011). *Wealth gaps rise to record highs between whites, blacks, Hispanics*. Pew Research Center.

Pollan, M. (2006). *The omnivore's dilemma: A natural history of four meals*. Penguin Press.

Porter, J., & Jick, H. (1980). Addiction rare in patients treated with narcotics. *New England Journal of Medicine*, 302(2), 123.

Rabin, R. (1989). Warnings unheeded: A history of child lead poisoning. *American Journal of Public Health*, 79(12), 1668–1674.

Rakoff, J. S. (2014, November 20). The financial crisis: Why have no high-level executives been prosecuted? *The New York Review of Books*.

Reuters. (2025, November 24). Meta shut down internal research after finding causal link to mental health harms, court filings say. *Reuters*.

Roose, K. (2025, April 24). If A.I. systems become conscious, should they have rights? *The New York Times*.

Rugh, J. S., & Massey, D. S. (2010). Racial segregation and the American foreclosure crisis. *American Sociological Review*, 75(5), 629–651.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Savitt, T. L. (2007). *Race and medicine in nineteenth- and early-twentieth-century America*. Kent State University Press.

Sen, A. (1981). *Poverty and famines: An essay on entitlement and deprivation*. Oxford University Press.

Shiva, V. (1991). *The violence of the Green Revolution: Third World agriculture, ecology, and politics*. Zed Books.

Starr, P. (1982). *The social transformation of American medicine*. Basic Books.

Twenge, J. M. (2017). *iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy, and completely unprepared for adulthood*. Atria Books.

United Nations Human Rights Council. (2018). *Report of the independent international fact-finding mission on Myanmar* (A/HRC/39/64). United Nations.

United States v. The Purdue Frederick Company, Inc., No. 1:07-cr-00029 (W.D. Va. May 10, 2007).

U.S. Environmental Protection Agency. (2021). *Addressing uranium contamination on the Navajo Nation*. EPA.

U.S. Senate Finance Committee. (2021). *Insulin: Examining the factors driving the rising cost of a century-old drug*. U.S. Senate.

Van Zee, A. (2009). The promotion and marketing of OxyContin: Commercial triumph, public health tragedy. *American Journal of Public Health*, 99(2), 221–227.

Wells, G., Horwitz, J., & Seetharaman, D. (2021, September 14). Facebook knows Instagram is toxic for teen girls, company documents show. *The Wall Street Journal*.

Williams, D. R., & Collins, C. (2001). Racial residential segregation: A fundamental cause of racial disparities in health. *Public Health Reports*, 116(5), 404–416.

Wu, T. (2016). *The attention merchants: The epic scramble to get inside our heads*. Alfred A. Knopf.

Zahran, S., McElmurry, S. P., Kilgore, P. E., Mushinski, D., Press, J., Love, N. G., Sadler, R. C., & Swanson, M. S. (2018). Assessment of the Legionnaires' disease outbreak in Flint, Michigan. *Proceedings of the National Academy of Sciences*, 115(8), E1730–E1739.

Zakrzewski, C. (2025, September 8). Meta suppressed children's safety research, whistleblowers tell Congress. *The Washington Post*.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.