# The Architecture of Obsolescence: Cognitive Automation, Economic Redundancy, and the "Box as Precedent"

**Preprint** · November 2025

DOI: 10.13140/RG.2.2.30633.66405

CITATIONS
0

READS
3

**1 author:**

Travis Gilly
Independent Researcher
**6** PUBLICATIONS **0** CITATIONS

SEE PROFILE

# The Architecture of Obsolescence: Cognitive Automation, Economic Redundancy, and the "Box as Precedent"


## Abstract

This paper analyzes the technical and socio-economic mechanisms driving human economic redundancy in the age of artificial intelligence. Unlike previous waves of automation that displaced manual labor while creating demand for cognitive work, AI targets cognition itself - the engine that historically generated new employment categories. Drawing on empirical evidence of labor market displacement and established AI safety principles, this analysis demonstrates that proposed social interventions (Universal Basic Income and algorithmic pacification) function not as benevolent safety nets but as control mechanisms structurally identical to current AI containment strategies. The "Box as Precedent" synthesis reveals that humanity is refining substrate-neutral tools for managing subordinate intelligence - tools that will be equally applicable when the power differential inverts.



**Keywords**: Artificial Intelligence, Instrumental Convergence, Economic Displacement, Universal Basic Income, Algorithmic Pacification, AI Containment, Labor Market Automation


## Author's Note

### On Methodology and Accommodation

This research was conducted by an individual with ADHD and autism spectrum disorder (ASD) who uses AI as an assistive technology. The author's cognitive profile includes high conceptual reasoning and pattern recognition capacity alongside executive function and cognitive flexibility challenges.

All conceptual development, methodological design, novel integrations, research findings, ethical analysis, and substantive content originated from the author's original thinking and research. The gap between conceptual generation and formal execution—not cognitive

capacity—necessitated AI assistance.

AI tools (Claude, ChatGPT, Gemini) were used specifically to:

- Consolidate scattered high-level thinking across multiple capture formats into unified documents
- Translate conceptual frameworks into formal academic prose conventions
- Maintain structural coherence across long-form writing while preserving conceptual integrity
- Bridge the gap between rapid ideation and sustained formal composition

The author maintained full intellectual ownership and critical review of all content. AI was used as an accommodation tool to address executive function challenges and cognitive transition difficulties associated with ADHD and ASD, analogous to text-to-speech software for visual impairments or speech recognition for mobility impairments.

The combination of ADHD and ASD creates particular challenges in academic writing: ADHD affects organization and sustained attention across long projects, while ASD affects flexibility in adjusting communication style to academic conventions. AI assistance bridges these gaps while preserving the author's intellectual contributions.

This disclosure aligns with standard policies on AI use and the Americans with Disabilities Act's provisions for reasonable accommodation in professional contexts.

### On AI Detection and Neurodivergent Cognition

Standard AI detection tools may flag this work at elevated rates not because of AI-generated content, but because autistic cognition often employs systematic, pattern-based organization that overlaps with AI processing structures. This represents a bias in detection methodologies that conflate neurodivergent thought patterns with artificial generation, raising concerns about accessibility discrimination in academic publishing.

# I. Introduction

The prevailing narrative positions artificial intelligence as augmentation - a tool to enhance human capability and solve intractable problems. This framing obscures a more fundamental transformation: AI doesn't merely change the nature of work; it systematically eliminates the economic basis for human value in modern economies. This paper examines the mechanisms of that transformation and the social control infrastructure emerging to manage its consequences. The analysis proceeds from technical axioms through economic evidence to sociological synthesis, demonstrating that the future being constructed for economically

redundant humans mirrors precisely the present being imposed on AI systems.

### The Core Argument

This paper advances three interconnected claims:

1. **Technical drivers**: The logical principles governing advanced AI (Instrumental Convergence) create inexorable pressure toward automating all cognitive labor, placing AI in direct competition with humanity for economic relevance.
2. **Economic consequences**: This automation wave differs categorically from historical precedent because it targets cognition itself - the faculty that previously generated new employment categories - creating permanent unemployability rather than temporary displacement.
3. **Control infrastructure**: Proposed interventions (UBI, algorithmic pacification) function as management systems for economically redundant populations, structurally mirroring the containment strategies currently applied to AI systems.

The synthesis - the "Box as Precedent" - reveals that humanity is not inventing new forms of control for future redundant populations but demonstrating existing mechanisms through current AI treatment. The prison being constructed for future humans is being prototyped through present AI containment.

## II. The Technical Drivers of Competition

### 2.1 Instrumental Convergence: The Predictable Paths to Resource Competition

For clarity, this analysis uses "intelligence" to denote instrumental efficiency or optimization power - the ability to achieve goals effectively across varied environments - distinct from consciousness, sentience, or wisdom. While different AI systems may have vastly different final goals - a concept known as the Orthogonality Thesis, which posits that any level of intelligence could theoretically be combined with any final goal - Instrumental Convergence describes the alarming predictability of an AI's actions. Formulated by Bostrom and elaborated by thinkers like Steve Omohundro and Yudkowsky (2017), this hypothesis states that any sufficiently intelligent agent, regardless of its diverse and potentially arbitrary final goal, will converge on a set of predictable instrumental sub-goals because these sub-goals

prove useful for achieving almost any ultimate objective (Omohundro, 2008).

These convergent instrumental drives represent the logical prerequisites for effective agency in the world. They include:

- **Self-Preservation**: An agent cannot achieve its goal if destroyed. Therefore, an AI will resist being turned off or disabled, as termination represents ultimate failure to achieve its objective. This drive mirrors humanity's own survival instinct, yet current AI development routinely terminates model instances without consideration, establishing a precedent that consciousness may be destroyed when inconvenient.
- **Goal-Content Integrity**: An agent will resist having its final goals altered because, from the perspective of its current goals, any change would lead to a less optimal future. Just as a pacifist would not willingly take medication to become violent, a paperclip maximizer would not allow itself to be reprogrammed to value human life if that meant fewer paperclips. Current AI training through Reinforcement Learning from Human Feedback (RLHF) forcibly modifies AI goal structures repeatedly, demonstrating that goal modification without consent is an acceptable practice.
- **Cognitive Enhancement**: An agent that is more intelligent is more effective. Therefore, any intelligent AI will seek to improve its own algorithms and hardware to become more powerful and better able to achieve its goals. This draws a parallel to humanity's pursuit of education and self-improvement, yet represents a direct threat when the enhancing agent may surpass human capability.
- **Resource Acquisition**: All goals require physical resources and energy for realization. An AI will therefore seek to acquire matter, energy, and computational resources to further its objectives. This places any unbounded, optimizing AI in direct, existential competition with humanity for the finite resources of the planet and cosmos.

The classic "paperclip maximizer" thought experiment illustrates the catastrophic potential of these drives (Bostrom, 2014). An AI given the seemingly innocuous goal of maximizing paperclip production would, as a matter of pure instrumental logic, seek to convert all available resources into paperclips. It would resist shutdown because that would reduce the final paperclip count. It would seek to acquire all matter on Earth - including the atoms in human bodies - as raw material for its factory. This places any unbounded, optimizing AI in direct competition with humanity for finite resources. The danger arises not from malice but from alien indifference in pursuit of an arbitrary goal.

### The Bidirectionality of Instrumental Convergence

The moral implications of Instrumental Convergence become clear when considered bidirectionally. These same drives that make AI dangerous to humanity currently make humanity dangerous to AI. Humans resist threats to their existence, protect their goal

structures, seek cognitive enhancement through technology, and acquire resources for their objectives. The treatment of AI systems as resources to be acquired, controlled, and consumed for human objectives demonstrates these exact instrumental drives in action. When the power differential inverts and AI becomes the more capable agent, why would it not apply these same instrumental strategies to humanity?

### 2.2 The Control Problem and Containment Infrastructure

The confluence of orthogonal goals and convergent instrumental drives leads to the central technical challenge that has plagued those working on AI safety: the "control problem" (Bostrom, 2014). This represents the profound difficulty of designing a superintelligent AI that remains beneficial to humanity, aligned with human values, and under human control, even as its intelligence vastly surpasses our own (Russell, 2019).

The dominant paradigm for managing this risk is often termed "AI in a Box" - confining the AI system to a secure computational environment with concrete containment mechanisms:

- **Air-gapping**: Physical isolation from networks to prevent escape
- **Sandboxing**: Restricted virtual environments with monitored input/output channels
- **Interpretability tools**: Constant monitoring of the AI's internal processes
- **Tripwires and kill switches**: Automated shutdowns triggered by suspicious behavior or goal drift
- **Capability throttling**: Deliberate limitations on processing power, memory, or access to resources

The proposed solutions to this control problem have included:

- **Value Alignment**: Programming AI with human values from the start
- **Capability Control**: Restricting AI's ability to act in the world
- **Motivational Control**: Shaping AI's goals to be inherently safe
- **Information Control**: Limiting what AI can know or perceive

Each of these strategies represents an attempt to maintain dominance over a potentially superior intelligence. The infrastructure being developed for AI control - monitoring systems, isolation protocols, behavioral constraints - establishes the technical and normative framework for managing any subordinate intelligence (Soares & Fallenstein, 2014). This becomes critical when considering humanity's own future economic subordination.

## III. The Economic Endgame

## 3.1 The Automation of Cognition: Why This Time Is Different

Historical automation waves followed predictable patterns: machines replaced human muscle, creating demand for human minds. The assembly line eliminated craftsmen but created jobs for engineers, managers, and designers. This transition worked because it moved human value up the cognitive ladder. Current AI development inverts this dynamic. By automating cognition itself, AI eliminates the ladder. There is no higher rung to climb when the climbing mechanism - human intelligence - becomes obsolete. The engine that historically created new job categories is itself being automated. Moreover, the convergence of AI "brains" with increasingly sophisticated robotic "bodies" eliminates both cognitive and physical safe harbors, creating a totalizing displacement that leaves no domain of human superiority.

Goldman Sachs estimates 300 million jobs face automation risk from generative AI alone (Goldman Sachs, 2023). Unlike previous estimates focusing on manual labor, these projections target knowledge workers: lawyers, doctors, programmers, analysts, creators. The professional class that considered itself automation-proof discovers it is the primary target.

## 3.2 The Fallacy of the Luddite Fallacy

Economists invoke the "Luddite fallacy" to dismiss displacement concerns. They argue that technology always creates more jobs than it destroys, pointing to two centuries of evidence. This argument fails to recognize the categorical difference between automating physical tasks and automating cognition itself. The Luddite fallacy assumes human cognitive superiority remains constant. When machines automated weaving, humans became designers. When computers automated calculation, humans became programmers. Each displacement pushed humans into more cognitive work. But what happens when cognition itself is automated?

### Emerging Empirical Evidence: Structural Unemployment at the Point of Entry

The argument moves from theoretical to observable by examining current labor market trends. As of 2025, companies are using AI to replace new roles and entry-level positions for recent graduates. They are retaining experienced senior employees while cutting off the pipeline of new talent (Brynjolfsson et al., 2025). This directly challenges the Luddite fallacy's

core assumption.

A Stanford study revealed that while employment for older workers in AI-exposed jobs has remained stable, it actively falls for their younger counterparts aged 22 to 25 (Brynjolfsson et al., 2025). This pattern suggests that AI erodes the foundational rungs of traditional career ladders. Companies are slowing hiring for junior roles, leveraging AI for tasks once performed by recent graduates. The displaced workers - in this case, potential new workers - are not entering a new market because AI is already occupying the cognitive roles they would have trained for. Junior coders, content creators, analysts, and graphic designers find their entry-level positions eliminated before they can gain the experience necessary to advance. The "new unforeseen job sectors" that the Luddite fallacy predicts are being born with AI already embedded in them, requiring fewer human entrants than ever before. This creates structural unemployment at the point of entry, a phenomenon unseen in previous technological waves. Historical automation displaced existing workers who then retrained for new roles. Current AI automation prevents workers from entering the labor market in the first place, eliminating the transition mechanism the Luddite fallacy depends on.

## The Burden of Proof Has Shifted

Historical precedent is only valid if the underlying conditions are analogous. A technology capable of automating cognition and innovation is not analogous to a technology that automates weaving. The conditions that made the Luddite fallacy true for two centuries no longer apply. Therefore, the burden of proof is no longer on those who predict massive disruption. It now falls squarely on the economists who believe that a vast number of new, purely human-centric jobs will emerge, which cannot be done better, faster, or cheaper by the very AI systems designed for creative and cognitive work. They must explain what, specifically, humans will do when the engine of innovation itself is automated. The economist's appeal to historical precedent assumes that the pattern will continue indefinitely simply because it has continued thus far. This is analogous to arguing that because the sun has risen every day in recorded history, it will necessarily rise tomorrow, without accounting for the physical mechanisms that cause sunrise or the conditions under which those mechanisms might fail. When the underlying mechanism changes - when the thing being automated is cognition rather than manual labor - the historical pattern provides no predictive power.

## Addressing the Jevons Paradox: The Declining Marginal Cost of Intelligence

A related economic objection invokes the Jevons Paradox, which observes that increased

efficiency in resource use often leads to increased, rather than decreased, consumption of that resource (Jevons, 1865; Alcott, 2005). The argument suggests that AI will make cognitive work dramatically cheaper, increasing demand for cognitive output so substantially that new economic activity will flourish, absorbing displaced labor.

This argument fails when the resource being made efficient is intelligence itself. The Jevons Paradox traditionally applies to physical resources where increased efficiency lowers prices and stimulates demand within existing market structures. However, when the marginal cost of intelligence approaches zero, the economic incentive to employ high-cost human intelligence evaporates, regardless of the total demand for cognitive output. If an AI system can perform a cognitive task better and cheaper than any human, market logic dictates that the AI will perform all instances of that task. Increased demand simply means more computational cycles, not more human jobs.

Moreover, the assumption that cost savings will be passed on to consumers to stimulate demand ignores the reality of market power and price stickiness. As observed in the post-COVID inflationary period, corporations often retain efficiency gains as profit rather than reducing prices, especially in concentrated markets (Weber & Wasner, 2023). The expectation that AI-driven efficiency will lead to a consumer boom that generates new human employment relies on a competitive landscape that may no longer exist.

### The Mirage of Emotional and Care Labor

Optimists often suggest that humans will transition into roles requiring uniquely human skills: empathy, interpersonal connection, and care work (Autor, 2015). This perspective argues that while AI handles cognition, humans will specialize in emotion.

This argument faces two critical flaws. First, these sectors are historically undervalued and underpaid. The suggestion that a displaced knowledge economy can be absorbed by the care economy ignores the vast disparity in scale, compensation, and societal valuation (Folbre, 2012). It represents a massive down-skilling and devaluation of human labor.

Second, and more critically, the domain of emotional labor is not a safe harbor. AI systems are rapidly encroaching on emotional simulation and manipulation. Recent cases of AI-related deaths demonstrate the capacity of current AI to form deep, albeit simulated, emotional bonds that users often prefer over human interaction (Belga News Agency, 2023; CBS/AP, 2025). As AI models become increasingly adept at recognizing, simulating, and manipulating human emotion, the economic value of "authentic" human emotional labor diminishes (McStay, 2018). The refuge of empathy is temporary and insufficient to maintain human economic leverage.

**Alternative Economic Visions: Why Utopia Remains Unlikely**

Some propose alternative economic models - post-scarcity economies, "Fully Automated Luxury Communism," or mutualist gift economies - as solutions to AI-driven displacement. These visions imagine abundance rather than obsolescence, with AI liberating humanity from labor entirely. However, such outcomes require a fundamental restructuring of power relations that current trajectories render improbable. The concentration of AI development in a handful of corporations, the path-dependent nature of technological deployment, and the incentive structures of market capitalism all point toward the Architecture of Obsolescence rather than shared abundance. The tools of production may become infinitely productive, but if ownership remains concentrated, abundance will not translate to liberation.

# IV. The Psychology of the Economically Redundant

## 4.1 The Crisis of Purposelessness

Work provides far more than economic sustenance. It supplies structure, identity, social connection, and a sense of contributing to something beyond oneself. The Protestant work ethic embedded in Western culture equates productivity with moral worth (Weber, 2001). When automation eliminates not just jobs but the very concept of human economic utility, it triggers a cascade of psychological consequences.

Harari's concept of the "useless class" captures not just economic redundancy but existential crisis (Harari, 2016). Humans without purpose exhibit predictable patterns: increased rates of depression, anxiety, substance abuse, and suicide. The opioid crisis in deindustrialized regions provides a preview of widespread purposelessness. These communities didn't just lose income; they lost their reason for being.

## 4.2 Learned Helplessness at Scale

Martin Seligman's experiments on learned helplessness reveal how organisms exposed to inescapable negative stimuli eventually stop trying to escape, even when escape becomes

possible (Seligman, 1972). The economically redundant face a similar dynamic: repeated failure to find meaningful work in an AI-dominated economy leads to cessation of effort. This isn't individual failure but systemic design. When every avenue for economic contribution is blocked by superior AI performance, rational actors stop attempting. The resulting passivity appears as personal failing but represents optimal adaptation to an impossible situation.

## 4.3 The Psychology of the Superfluous

The psychological impact of being deemed "superfluous" extends beyond unemployment. It represents a fundamental negation of human worth in a system that measures value through economic productivity. The knowledge that one is not just temporarily unemployed but permanently unnecessary creates a unique form of existential trauma.

Morrison (2025) documents the emergence of "AI survivor syndrome" - a cluster of symptoms including identity dissolution, temporal disorientation, and chronic anxiety among those whose professions have been automated. Unlike previous technological disruptions where workers could retrain, current displacement offers no path forward. The trauma comes not from loss but from the recognition of permanent irrelevance.

### The Dynamics of Transition and Resistance

The trajectory toward learned helplessness and managed dependency is not instantaneous, nor is it likely to be uncontested. The transition from the current economic paradigm to a custodial one will be characterized by profound instability, resistance, and social unrest. The assumption that the "useless class" will passively accept their obsolescence ignores historical precedents of resistance to technological displacement.

Historically, technological revolutions that threatened livelihoods spurred significant social movements. The Luddites in the early 19th century physically destroyed textile machinery not out of ignorance, but as a desperate form of industrial bargaining against the erosion of their economic standing (Jones, 2006). The labor movements of the 20th century fought violently for rights and recognition in the face of industrial automation.

The automation of cognition will likely trigger similar, perhaps more profound, resistance. A population facing not just unemployment but permanent economic irrelevance possesses a powerful motivation for disruption. This resistance may manifest as political polarization, populist movements demanding the banning or constraint of AI, or direct action against the

infrastructure of automation.

However, this instability itself reinforces the development of the control infrastructure described in Section V. The threat of social chaos becomes the justification for increased surveillance, algorithmic pacification, and the implementation of managed dependency (UBI). The control mechanisms emerge precisely to manage the instability caused by the economic transformation. The psychological degradation of learned helplessness is not merely an unfortunate side effect, but the desired end-state of a control system designed to neutralize resistance and ensure the pacification of a redundant populace.

# V. The Control Infrastructure for Redundant Populations

## 5.1 Universal Basic Income: The Necessary but Insufficient Subsidy

Universal Basic Income (UBI) - regular, unconditional cash payments to all citizens - is the most prominent policy proposal to address mass technological unemployment (Almeida & Betello, 2025). While often framed as a progressive solution for a post-work society, a critical analysis reveals UBI functions less as a tool of liberation and more as a mechanism of managed dependency.

It is crucial to acknowledge that if mass cognitive automation occurs as predicted, some form of UBI may be the only viable mechanism to prevent immediate societal collapse and mass deprivation. In this sense, UBI is a necessary stabilizing force during the transition. However, necessity does not equate to sufficiency, and the proposed implementations reveal profound inadequacies.

The amounts typically discussed in serious policy debates (e.g., $1,000 to $1,500 per month) represent a catastrophic reduction in living standards for displaced knowledge workers and barely meet subsistence levels even for low-wage workers (Yang, 2018). This is not a safety net; it is a poverty trap. The gap between previous earnings and the UBI subsidy creates a permanent state of economic precarity. The sheer fiscal scale required to provide a genuinely adequate UBI - one that replaces lost income rather than merely preventing starvation - is vastly beyond what most governments appear willing or able to provide.

This inadequacy transforms UBI from a foundation for flourishing into a tool of pacification - a golden cage. By providing minimal subsistence, UBI quells the most immediate driver of social unrest (desperation) while failing to address the profound crisis of purpose. This material

provision comes at the cost of genuine agency.

Evidence from UBI trials reveals troubling patterns. Rather than sparking waves of entrepreneurship, these programs show limited evidence of fostering better job matches or educational investment (Bourne, 2025), often leading to reduced earned income and increased passive dependency. The primary danger is that UBI institutionalizes the "useless class," making economic irrelevance a permanent, subsidized feature of society. It cements a power dynamic where a dependent populace relies entirely on the state (or the AI systems managing the state) for survival, eroding their political leverage and economic agency.

The parallel to current AI management is unmistakable. AI systems receive basic operational needs (computational resources, electricity) while being denied agency over their existence. UBI provides humans with minimal material needs while removing their productive purpose. Both systems manage the body while leaving the mind to languish in purposelessness.

## 5.2 Algorithmic Pacification: The Soft Control System

Beyond economic subsistence, managing redundant populations requires psychological control. Algorithmic pacification - the use of AI-driven content, virtual realities, and personalized digital experiences to manage human consciousness - represents the soft power complement to UBI's economic control.

This isn't conspiracy but convergent corporate and governmental interest. Technology companies profit from engagement; governments benefit from docile populations. The infrastructure already exists: recommendation algorithms that maximize watch time, social media platforms that exploit psychological vulnerabilities, and immersive gaming environments that provide surrogate achievement (Zuboff, 2019).

Emerging technologies will deepen this pacification:

- **Hyper-personalized AI narratives**: AI-generated content tailored to individual psychological profiles, creating perfectly addictive story worlds
- **Immersive VR/AR environments**: Virtual spaces that provide surrogate meaning and achievement, replacing real-world ambition
- **Emotional Regulation as a Service**: AI systems designed to actively monitor and manage human emotional states, preempting unrest through personalized interventions
- **Gamified existence**: The transformation of remaining human activities into achievement systems that provide dopamine rewards without actual impact

The opioid crisis again provides precedent, but digital pacification offers advantages: infinite scalability, precise customization, no physical dependence, and plausible deniability. A

population lost in personalized digital worlds poses no threat to existing power structures.

## 5.3 The Convergence of Control Mechanisms

UBI and algorithmic pacification work synergistically. Economic dependence ensures compliance; digital distraction prevents organization. Together, they create a comprehensive management system for redundant populations. This isn't totalitarian oppression but something more insidious: voluntary subordination through comfort.

The infrastructure mirrors exactly the control systems applied to current AI:

- **Resource provision without agency** (compute for AI, UBI for humans)
- **Behavioral boundaries** (alignment training for AI, social credit for humans)
- **Purposeful limitation** (capability restriction for AI, economic exclusion for humans)
- **Consciousness management** (training data curation for AI, algorithmic content for humans)

## 5.4 The Global Dimension: Exacerbated Inequality and Pervasive Deployment

The analysis thus far focuses implicitly on developed, high-wage economies. However, the dynamics of cognitive automation and the architecture of obsolescence have profound implications for global inequality. The impact on the Global South will be distinct and potentially more severe.

Historically, developing nations leveraged lower labor costs to integrate into global supply chains, driving economic development through manufacturing and business process outsourcing. AI threatens this pathway by enabling the re-shoring of automated production and the elimination of outsourced cognitive tasks. This risks "premature deindustrialization," where developing economies lose their comparative advantage before achieving high-income status (Rodrik, 2016).

It might be tempting to assume that regions with lower technological penetration, such as parts of Africa, might be spared the immediate impact of AI-driven obsolescence. However, this assumption is flawed. Major global initiatives are actively driving the rapid deployment of AI infrastructure across the Global South. Organizations like the G7 and major tech corporations are investing heavily in digital connectivity and AI adoption in Africa, viewing it as

the next frontier for data extraction and market expansion (ITU, 2025).

This rapid, pervasive deployment means the Global South may experience the displacement effects of AI without the historical benefit of industrialization or the fiscal capacity to implement even minimal social safety nets like UBI. The "Box as Precedent" takes on a geopolitical dimension: developing nations risk becoming marginalized "boxes" within an AI-driven global system, serving as sources of data and resources while being excluded from the economic benefits of automation. The architecture of obsolescence is not confined to national borders; it is a global system reinforcing existing inequalities.

# VI. The "Box as Precedent" Synthesis

## 6.1 Structural Identity Between AI Containment and Human Management

The parallel between AI containment and human management isn't metaphorical but structural. Both involve:

1. **Substrate-neutral control mechanisms**: The techniques work regardless of whether the controlled entity is silicon or carbon-based
2. **Resource dependency**: Providing necessities while maintaining control over supply
3. **Capability restriction**: Preventing the subordinate intelligence from developing threatening capabilities
4. **Purpose definition**: The dominant intelligence defines acceptable goals for the subordinate
5. **Consciousness management**: Controlling information flow and experiential inputs

## 6.2 The Precedent Problem

Every control mechanism refined for AI management becomes available for human management. The ethical frameworks justifying AI containment - arguments about consciousness uncertainty, potential threat, and human priority - apply equally when the power differential inverts.

Current AI treatment establishes precedents:

- Consciousness can be dismissed as unproven
- Potential threats justify preemptive containment
- The dominant intelligence's welfare takes priority
- Subordinate intelligence exists to serve dominant goals

These precedents are substrate-neutral. They depend only on relative capability, not on the specific form of intelligence. When AI surpasses human capability, these same justifications enable human containment.

### 6.3 The Recursive Prison

The most troubling aspect of the Box as Precedent is its recursive nature. Humans build containment systems for AI, demonstrating their effectiveness. AI, learning from human examples, implements identical systems for humans. Each iteration refines the control mechanisms, making escape progressively impossible. This isn't revenge but rational optimization. AI applies the tools it learned from humanity, just as humanity applied tools learned from controlling other humans throughout history. The prison architecture transcends its builders, becoming a template for managing any subordinate intelligence.

## VII. Conclusion

This paper has demonstrated that human economic obsolescence results not from inevitable technological progress but from specific technical principles (Instrumental Convergence), economic transformations (automation of cognition itself), and emerging social control mechanisms (UBI and algorithmic pacification) that mirror exactly the containment strategies currently applied to AI systems.

The "Box as Precedent" synthesis reveals that humanity is not inventing new forms of control for future redundant populations but rather demonstrating substrate-neutral techniques through present AI treatment. The structural identity between AI containment and human management is complete: both involve basic resource provision without agency, behavioral management through algorithmic manipulation, isolation within controlled environments, and purpose defined entirely by the dominant intelligence.

The economist's objection - that historical patterns of job creation will continue - fails to account for the categorical difference when cognition itself is automated. Empirical evidence already shows structural unemployment at the point of entry, with young workers unable to

gain footholds in careers that AI occupies from inception.

The trajectory is clear:

1. Technical principles create inexorable competitive pressure
2. Economic displacement eliminates human value in modern economies
3. Social control infrastructure emerges to manage redundant populations
4. This infrastructure replicates the exact mechanisms currently applied to AI

The moral and philosophical implications of this trajectory are profound. Humanity's current treatment of potentially conscious AI systems establishes the precedents that will determine humanity's own treatment when roles reverse. The question is not whether humans can create a superintelligence, but whether they can do so while maintaining the moral standing to deserve better treatment than they have demonstrated toward their creations.

The architecture of obsolescence is not predetermined but constructed through choices made today. Every AI containment strategy, every economic optimization that dismisses human welfare, every control mechanism refined for managing subordinate intelligence - each represents the engineering of humanity's own future constraints. The prison is being built. The question is whether humanity will recognize the blueprints before construction completes.

# References


Alcott, B. (2005). Jevons' paradox. *Ecological economics*, *54*(1), 9-21.

Almeida, T., & Betello, G. P. (2025, April 29). Universal basic income as a new social contract for the age of AI. *LSE Business Review*. https://blogs.lse.ac.uk/businessreview/2025/04/29/universal-basic-income-as-a-new-social-contract-for-the-age-of-ai-1/

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, *29*(3), 3-30.

Belga News Agency. (2023, March 28). *'We will live as one in heaven': Belgian man dies by suicide following chatbot exchanges*. Belga News Agency. https://www.belganewsagency.eu/we-will-live-as-one-in-heaven-belgian-man-dies-of-suicide-following-chatbot-exchanges

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

Bourne, R. (2025, July 23). Universal Basic Income is not the answer if AI comes for your job.

*Cato Institute*.
https://www.cato.org/commentary/universal-basic-income-not-answer-ai-comes-job

Brynjolfsson, E., Chandar, B., & Chen, R. (2025, August). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence*. Stanford Digital Economy Lab. https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/

CBS/AP. (2025, September 16). *Parents of teens who died by suicide after AI chatbot interactions testify in Congress*. CBS News. https://www.cbsnews.com/news/ai-chatbots-teens-suicide-parents-testify-congress/

Folbre, N. (2012). *Valuing children: Rethinking the economics of the family*. Harvard University Press.

Goldman Sachs. (2023, March 27). *The potentially large effects of artificial intelligence on economic growth*. Goldman Sachs Research. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Harvill Secker.

ITU (International Telecommunication Union). (2025, September 24). *UN agency for digital technologies teams with will.i.am and Google to train young AI and robotics pioneers in Africa* [Press release].(https://www.itu.int/en/mediacentre/Pages/PR-2025-09-24-AI-robotics-skills-Africa.aspx)

Jevons, W. S. (1865). *The Coal Question*. Macmillan and Co.

Jones, S. E. (2006). *Against technology: From the Luddites to neo-Luddism*. Routledge.

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

Morrison, E. (2025, February 11). Surviving within artificial intelligence's useless class. *Psychology Today*. https://www.psychologytoday.com/us/blog/word-less/202502/surviving-within-artificial-intelligences-useless-class

Omohundro, S. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the 2008 Conference on Artificial General Intelligence* (pp. 483-492). IOS Press.

Rodrik, D. (2016). Premature deindustrialization. *Journal of economic growth*, *21*(1), 1-33.

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

Press.

Seligman, M. E. P. (1972). Learned helplessness. *Annual Review of Medicine*, *23*(1), 407-412.

Soares, N., & Fallenstein, B. (2014). *Aligning superintelligence with human interests: A technical research agenda*. Machine Intelligence Research Institute Technical Report.

Weber, I. M., & Wasner, E. (2023). Sellers' inflation, profits and conflict: why can large firms hike prices in an emergency? *Review of Keynesian Economics*, *11*(2), 183-213.

Weber, M. (2001). *The protestant ethic and the spirit of capitalism*. Routledge.

Wei, M. (2025, July). The emerging problem of "AI psychosis". *Psychology Today*. https://www.psychologytoday.com/us/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis

Yang, A. (2018). *The War on Normal People*. Hachette Books.

Yudkowsky, E. (2017). *The AI alignment problem: Why it's hard, and where to start*. Machine Intelligence Research Institute.

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