JASC
Journal of AI, Society, and Critique

Volume 1, Number 1 (July 2026)

ASC
Center for AI, Society, and Critique

Journal of AI, Society, and Critique (JASC)

An international, open-access, peer-reviewed journal

ISSN (Online): 3056-056X

Editorial Information

Aims and Scope

The *Journal of AI, Society, and Critique (JASC)* is an international peer-reviewed journal dedicated to critical examinations of artificial intelligence from philosophical, sociological, and interdisciplinary perspectives, with particular attention to its social, cultural, political, and ethical implications.

Publication Information

Publisher: Center for AI, Society, and Critique (CASC)
Publication Model: Online, open access, peer-reviewed journal
Journal Homepage: https://aisocietycritique.org/journal/
Contact: info@aisocietycritique.org

Contents

Critical Theory of AI: The Alienation of the Human Intellect

Geoff M. Boucher[1]

Abstract: In this article, I present some fundamentals of a critique of AI influenced by the Critical Theory of the Frankfurt School. From this perspective, artificial intelligence represents a kind of apotheosis of instrumental rationality, an alienation of the human intellect, framed by economic calculations, geopolitical strategies, the reduction of reason to cognition, perception to classification and knowledge to an inventory. The article begins by defining machine-learning-based forms of artificial intelligence, distinguishing between different types of foundation model (LLM-based and non-LLM-based), describing how LLM-based GenAI works and discussing the realistic prospects for the emergence of AGI in the near future, especially when this is anticipated from increases in compute scale of GenAI. Then, it outlines the fundamental critical position of Frankfurt School Critical Theory, which is that human knowledge cannot be reduced to an inventory of items any more than the human intellect can be reduced to instrumental rationality. Against this background, it critiques the application of AI as a business machine and a warfare technology, arguing that it is crucial, in opposing corporate and military applications, not to uncritically accept the reduction of the human intellect to mental labor. Nonetheless, the

[1] **Geoff M. Boucher** is Associate Professor in Literary Studies at Deakin University. He currently researches the authoritarian personality from the perspective of critical theory and is an expert on Frankfurt School critical theory, Marxism, and psychoanalysis, as well as the work of Žižek and Lacan. He is the author of *Traversing the Fantasy: Critical Essays on Slavoj Žižek* (co-editor, 2005), *The Charmed Circle of Ideology* (Melbourne: Re-Press, 2008), *The Times Will Suit Them: Postmodern Conservatism in Australia* (with Matthew Sharpe) (Sydney: Allen & Unwin, 2008), *Žižek and Politics: A Critical Introduction* (with Matthew Sharpe) (London: Edinburgh University Press, 2010), *Adorno Reframed* (London: I.B. Tauris, 2012), *Understanding Marxism* (London: Acumen, 2012), *Habermas and Literature* (London: Bloomsbury Academic, 2021), and *Critical Theory and the Authoritarian Personality* (Edinburgh University Press, 2025).
In addition, he is the author of more than 40 scientific publications in literary criticism, cultural studies, and social theory, and has published numerous articles on right-wing authoritarianism. His particular interest in AI concerns the anti-democratic implications of misaligned or unregulated "grokbots."

implications of AI for the automation of mental labor are critically important, and so the article turns next to the microeconomics of AI implementation, inferring from the discourse of liberal economics that the disruptive structural transformation of the middle class through polarization and casualization is a real possibility. In conclusion, the article discusses the implications of this dialectical critique, not just in terms of the potential for a political convulsion, but also in terms of the urgent need for AI research to acknowledge (and address) the problems and limitations of current foundation models.

Introduction

Contemporary Frankfurt School Critical Theory is uniquely positioned to contribute to critical discourse on AI, because its humanist critique of instrumental rationality points right at the heart of the major limitation of the emergent machine-learning technological revolution. Artificial intelligence represents an alienation of the human intellect, framed by economic calculations, geopolitical strategies, the reduction of reason to cognition, perception to classification and knowledge to an inventory. By instrumental rationality, what is meant is a particularly useful but limited kind of reasoning which: (1) regards the world cognitively, as an object domain to be accurately and objectively reported; (2) in order to perform means-ends calculations oriented to the successful manipulation of these objects. As refined and developed by German philosopher Jürgen Habermas, the Frankfurt School critique of instrumental rationality is not a rejection of science (which emerges from cognitive reasoning) or technology (which emerges from instrumental manipulation) (Habermas, 1984: 273-337). It is first and foremost a critique of the runaway predominance of instrumental rationality, when this is indiscriminately implemented under conditions of corporate power, bureaucratic surveillance, state control and military competition (Habermas, 1987: 153-197). As a kind of apotheosis of instrumental rationality, current forms of machine-learning-based AI represent a narrow simulation of thinking that alienates the fullness and richness of the human intellect into success-oriented forms of token presentation and pattern recognition, deployed in the service of corporate profitability and superpower rivalry. Its irresponsible implementation in the context of deregulated markets is highly likely to cause deep social polarization, through a structural transformation of the middle class that may involve mass technological unemployment and occupational dislocation, possibly leading to a political convulsion reminiscent of the 1930s. Reckless corporate and government development of AI has the potential to

catalyze economic precariousness, cyber disruption, neo-fascist movements and war. It is crucial to recognize that despite its potential benefits for humanity, current forms of Large Language Model (LLM)-based Generative AI (GenAI) have not been built primarily for human well-being, but expressly for the purpose of replacing human mental labor with machine learning processes. OpenAI is explicit about what drives the entire industry, ever since at least the sea change in about 2023 from cautious and public AI research to reckless and private AI development (Hao, 2026). According to the makers of ChatGPT-5:

> "OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity" (OpenAI, 2026: online, emphasis added).

By Artificial General Intelligence (AGI), what is normally meant is an artificial form of the human intellect—the definition of general intelligence—often benchmarked against human measures of conceptual performance at post-PhD or Nobel Prize levels (as further detailed in the body of this paper). Visionary proponents of the technology, such as Anthropic CEO Dario Amodei, sometimes describe the possibilities of AGI (as normally defined) in terms of the potential benefits of having a nation of geniuses, housed in a datacenter, at the disposal of every user who wants to invent or discover something of benefit to humanity (Amodei, 2024). But in the OpenAI statement, which unfortunately is typical, not aberrant, this noble goal has been replaced by something entirely different, namely, a machine-learning technology that automates intellectual processes in a way similar to how machine action, such as the factory robot, automates manual labor. Although it is highly likely that the AI R&D corporations will triumphantly unveil their homunculus, "AGI™," at some point in the promised window—because they must—present indications are that it will fall far short of, say, passing Humanity's Final Exam, or even meeting functional competency scaling benchmarks such as "3-Expert" level performance. Instead, the aim of this highly political development of the technology is productivity improvement, through the expulsion of some workers from production (known as technological unemployment) and the imposition of speed-up on others (known as AI augmentation), something which the corporate world eagerly awaits as a profit

bonanza. In this vision of the technology, supposed improvements to human welfare follow from the capitalist cycle of "creative destruction," and in particular, from renewed profitability for multinational corporations and the improved position of superpower nations.

The likely social consequences of this development are already visible in the economics of AI diffusion. The CEOs of the AI R&D corporations are careful to warn that the workplace implementation of GenAI and its possible successor, AGI™, is expected to bring extraordinary levels of social suffering. Indeed, the manufacturers of this technology now implore governments to consider a basic income as an ameliorative measure designed to prevent a political convulsion. The gradual shift in media reporting from dismissal of the technology as "mid-tech" and liable to speculative collapse, to acknowledgement that the canaries in the coal mine are already falling and that this correlates with rising hostility to AI datacenters and technological unemployment, is instructive. It cannot be entirely ruled out that this technology might have a "Chernobyl moment," leading to its politically motivated abandonment as a business machine, and its retreat to a research specialization at the frontier between cybernetics and psychology. Certainly, the main indications that its commercial applications are a socially radioactive technology are well known even outside specialist circles—(1) its training data, potentially baked in to current foundation models, effectively silences the voices of the Global South while statistically normalizing racial and gender biases (Noble, 2018, Benjamin, 2019, Chun, 2024); (2) its environmental costs, the result of declining returns for scaling increases, based on linear computing improvements requiring exponential increases in energy demands, continue to worsen; (3) its unregulated development, combined with a lack of guardrails on its government abuse and private misuse, have seen intellectual property theft on a grand scale and the exploitation of training labor (O'Neil, 2017, Gray and Suri, 2019). This has been compounded by privacy violations, the metastasis of surveillance apparatuses, and the pollution of public debate by misinformation and denigration.

With the exception of Christian Fuchs and Jean-Philippe Deranty (whose positions are discussed in this paper), the leading contemporary thinkers of the Frankfurt School have yet to publish major statements on artificial intelligence. In the interim, this paper presents a provisional summary of the main points that follow reasonably straightforwardly from the approach of Critical Theory. The

critique that I propose builds on the theory of communicative action of Jürgen Habermas (Habermas, 1984, Habermas, 1987), the digital humanism of Christian Fuchs (Fuchs, 2022), Mark Coeckelbergh's politics of technology interpretation (Coeckelbergh, 2022), the critique of knowledge extractivism by Matteo Pasquinelli (Pasquinelli, 2024), the discussion of AI and labor by Jean-Philippe Deranty (Deranty and Corbin, 2024) and the micro-economics of AI proposed by Joshua Gans (Gans, 2026). It also builds on a very long and heavily referenced report that I am preparing on AI and Critical Theory for the Berlin-based Center for AI, Society and Critique, which explores the relevant scientific and technical literature in the AI R&D field in considerable detail. In what follows, I present a dialectical critique of the present path of AI development in forms of machine-learning-based AI, whether these take the current form of GenAI or the possible near-future form of AGI. As Fuchs writes, the critical position "rejects the idea of replacing humans by or transforming them into digital machines. Rather, it sees digital machines as possibilities that as part of struggles for a society that benefits all humans can expand, help realizing and more fully developing the potentials of humans and society" (Fuchs, 2022: 153). By critique, I mean a kind of criticism which goes to the roots of the thing, exploring its conditions of possibility in the simulation of thinking through instrumental models. Although it by no means fills up the device with some baleful essence, the instrumental nature of the technology predisposes it, in a society governed by the logic of functional systems (by capitalist economics and governmental bureaucracy), to development and implementation pathways that are inimical to human welfare. By dialectical, what I basically mean is a position which acknowledges that scientific advances and technological breakthroughs can be beneficial to humanity, but which is highly critical of the corporate and military implementation of AI. This dialectical approach is registered terminologically in the distinction (derived from Hegelian Marxism) between the externalization of human knowledge in a technological objectification, and its alienating effects when applied in opposition to human welfare, in the service of corporate profitability and imperial supremacy.

Artificial General Intelligence and GenAI

Although the development of Artificial General Intelligence (AGI) is the stated goal of major AI R&D corporations, such as OpenAI, Anthropic, Google, IBM, Microsoft and Meta, and of Chinese state AI strategy (Johnson, 2025: 14-20),

there is no consensus definition in the field about the performance benchmarks that would define when it has been achieved. Furthermore, as philosopher Nick Bostrom points out, the current industry-standard method, using stochastic processes within Large Language Models (LLMs) as the basis of machine learning (ML), is only one of three approaches to achieving cognitive performances on or beyond the present frontier of human intelligence (Bostrom, 2016: 27-54). Nonetheless, from the descriptive perspective, a commentary from IBM usefully states that:

> "Artificial general intelligence (AGI) is a hypothetical stage in the development of machine learning (ML) in which an artificial intelligence (AI) system can match or exceed the cognitive abilities of human beings across any task" (Bergmann and Stryker, 2026: 1).

In line with the original definitions that created the field (Goertzel and Pennachin, 2007: 1-2), the IBM commentary then specifies Artificial Narrow Intelligence (ANI) and Artificial General Intelligence (AGI), to which I add Nick Bostrom's definition of Artificial Superintelligence (ASI):[2]

- Artificial Narrow Intelligence (ANI): a ML-based AI that (significantly) exceeds human performance at specific tasks—such as playing Go or winning at chess;
- Artificial General Intelligence (AGI): a (hypothetical) ML-based AI that matches or exceeds human performance across all relevant cognitive domains;
- Artificial Superintelligence (ASI): a (speculative) entity that has "an intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2016: 26).

(Bostrom's discussion makes it clear that "greatly exceeds" should be grasped in terms of orders of magnitude.)

In what follows, I shall concentrate on the relation between AGI and Large Language Models (LLMs) running as Generative-AI (GenAI), such as is illustrated by ChatGPT5 (OpenAI), Claude (Anthropic), Google Gemini,

[2] The IBM definition conflates ANI and ASI because it defines superintelligence as superhuman task performance across one or more domains, rather than across every cognitive domain of relevance.

Microsoft Copilot, (Meta) Llama 4, DeepSeek, and so forth. First, I will discuss pathways from ANI and LLM-based GenAI to AGI. Second, I shall provide some more concrete markers for AGI by discussing proposed benchmarks. In the following section, I shall discuss the difference between "Strong-AI" and "Weak-AI" positions on the hypothetical technology. Before beginning, however, I should state categorically:

- There is no credible evidence that current GenAI meets any proposed performance standard for AGI (Tihanyi, Bisztray et al., 2024);
- There is no credible evidence that current GenAI exhibits self-consciousness, or even real awareness (Butlin, Long et al., 2023).

I now need to introduce some distinctions based on Rishi Bommasani's and Percy Liang's (and co-thinkers') seminal paper "On the Opportunities and Risks of Foundation Models". Intended to bridge the gap between ANI and AGI, Foundation Models are large-scale, pre-trained, general purpose models that are "trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks," through key processes of transfer learning and task adaptation (Bommasani, Liang et al., 2022: 1). These machines use stochastic processes to simulate inductive learning, leading to predictive simulations (e.g., protein folds), classificatory results (e.g., radiological applications) and generative textual or image responses (e.g., ChatGPT-5). GenAI refers to the subclass of Foundation Model that produces generative outputs as well as predictions and classifications. For present purposes, Foundation Models that are forms of GenAI fall into two categories, those that are Large Language Models (natural language based) and those that are not (e.g., DALL·E—now GPT Image-2). Generative forms of Foundation Model that are LLMs are what most readers will have experienced, in the form of Claude, ChatGPT, Gemini, Llama, DeepSeek and so forth. According to Bommasani and Liang, Foundation Models display properties of homogenization (there is a tendential convergence on a shared model architecture, often known as Transformer Architecture) and emergence (they exhibit emergent properties, such as generation from prompts, which are not trained responses) (Bommasani, Liang et al., 2022: 2-4). Because the foundational model provides the basis for subsequent developments, there is a significant risk of the inheritance of defects (e.g., biases in the training data, or biased training parameters), while emergence is the most important property for consideration of the potential originality of machine outputs.

Current versions of LLM-based GenAI use machine-learning processes applied to vast quantities of (mostly) human-generated text to identify statistical patterns in language (and other forms of signification), in order to generate—through prediction and transformation—human-intelligible, sentence-like outputs. Output text is produced by models trained to approximate conditional probability distributions over sequences of tokens, in a manner analogous (though not identical) to Bayesian inference, which is typically refined through training regimes that include human evaluation of model outputs. The results of this training process—compressed representations of probabilistic regularities and pattern interactions—are stored as billions of numerical parameters or "weights," each of which can be understood, in loose Bayesian terms, as resembling a posterior distribution collapsed into a single point estimate. Once training concludes, these weights are fixed, and the model applies its learned generative function to user inputs, articulated in natural language as "queries" that yield probabilistically generated "responses". These responses are dynamic and context-sensitive by virtue of Transformer Architectures, which represent tokens (effectively, though not exactly, semantic units) as vectors in multi-dimensional space, such that statistically related meanings are encoded as spatial proximity. Relevance is computed through attention mechanisms that assign dynamic, context-dependent weights to token–token relationships—effectively answering the question: "how much should token A matter to token B right now, given this context?" Contemporary LLMs stack multiple layers of such attention-based computation, enabling progressively higher-order patterning that moves from syntax through discourse, genre, rhetoric, and pragmatic inference, thereby producing outputs that mimic reasoning. In sum, most current LLMs combine a probabilistic objective (predicting sequences of response tokens), a parameterized function (its stored weights), and a transformer architecture (that performs attention-based, relevance-oriented calculations).

LLMs have been described, derisively but not absurdly, as stochastic parrots (Bender, Gebru et al., 2021). As this description implies, GenAI does not, strictly speaking, "think" about "problems" or provide "answers" to "questions". Rather, it probabilistically simulates these behaviors without engaging in reasoning as such. Nonetheless, speculation regarding an incremental pathway from GenAI to AGI has now become commonplace. There are solid grounds for

scepticism here, including models' performance against demanding benchmarks. These come in two types—final exams, and capability scaling.

Final Exam-type Proposals

Despite the enormous influence of Alan Turing's linguistic indistinguishability test (the "Turing Test") (Turing, 1950), the proposed benchmark ability of a machine to conversationally simulate a human interlocutor is not regarded as a satisfactory test instrument in AI research today (Levesque, 2014). Alternatives to the Turing Test include:

- Humanity's Last Exam (HLE): a 2,500-item test, mainly in STEM areas but with some humanities questions, "a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind" (Phan, Gatti et al., 2025: 4). The best three contenders score below 10% (i.e., "abject fail") (e.g., OpenAI o1: 8.8%; Gemini 2.0 Flash: 7.2%; Claude 3.5 Sonnet: 4.8%) (Maslej, Fattorini et al., 2025: 141-142).
- Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI-1, ARC-AGI-2 and ARC-AGI-3): the ARC-AGI proposal measures fluid intelligence in finding solutions to novel problems (Chollet, 2019: 2). Last year, the current set of LLMs scored in the 50% range on ARC-AGI-1 but only in the 3% range on the more advanced ARC-AGI-2 (Chollet, Knoop et al., 2025: 8-9). Recently, Claude Opus 5 (High) scored 97.5% on ARC-AGI-1, 90.4% on ARC-AGI-2 and 30.23% on ARC-AGI-3 (ARC-AGI-Prize, 2026, Chollet, Knoop et al., 2026).

Graduated Continuum Proposals

Rather than a final exam, some researchers have proposed that AGI is a spectrum of achievements on a continuum. Although no formal benchmark equates LLMs with specific educational stages, a widely-used informal vocabulary provides analogies to student performance, such as "a good high school student" (ChatGPT-3), "a good undergraduate student" (ChatGPT-4) and "a second-year postgraduate student" (ChatGPT-5). This is encouraged by the fact that OpenAI, for instance, benchmarks its technical reports against human educational cohorts (OpenAI (Achiam et al.) and Ecoffet (Corresponding Author), 2024).

Continuum proposals include

- Education-standardized Human-centered Tests (AGI-Eval): the AGI-Eval test instruments benchmark AI against human performance on university-level entrance exams and mathematics competitions, at levels of achievement well below "frontier benchmarks in human knowledge"—but well above median human performance. Here, LLMs such as ChatGPT-4 pass the SAT, sometimes scoring better than most human subjects on the LSAT, SAT, and SAT-Math tests (scores in the 60%-95%+ range depending on the test and whether it is with or without practice attempts) (Zhong, Cui et al., 2023: 10).
- Competency-relative Power Scaling: Based on six principles whose key is "focus on the path to AGI, not a single endpoint," a Google DeepMind team proposes a five-step scale in the human range: (1) emergent, (2) competent, (3) expert, (4) exceptional, and (5) superhuman. Competent AGI would have the same capacities as 50% of the human population, Expert AGI performs better than 90% of humans, Exceptional AGI performs in the 99% range and Super AGI outperforms 100% of human beings. While not an IQ test, the positioning of categories relative to population percentiles is suggestive; the team controversially suggests that current LLMs are at the "1-Emergent" (unskilled human) level (Ringel Morris, Sohl-Dickstein et al., 2025: 8).

What these benchmarks collectively demonstrate is that current LLMs are far from AGI by any demanding measure, with a gap that is not merely quantitative but qualitative.

Test Performances, Reasonable Expectations and Instrumental Rationality

The variation in approaches to what is tested and how this is benchmarked explains the uncertainty in predictions regarding the achievement of AGI. According to the website Our World in Data, 50% of expert predictions indicate a date before 2061 for complete AGI, with the median prediction about 2042 (that is, effectively, an AI passing a final exam) while 50% of Metaculus analysts predicted the release of a system described as AGI by 2040 (that is, effectively, an AI located at Competent or Expert on some scale) (Roser, 2023). According to the website AI Multiple, analysis of 9,800 predictions indicates recent strong convergence on 2031-2032 as the date for the achievement of AGI, defined as "matching human cognitive ability across all tasks" (that is, effectively, an AI passing an intermediate power scale test) (Dilmegani and Ermut, 2026). It is worth

asking "why?" there is such a discrepancy between the expected release of something called "AGI" and its passing a final exam for AGI. Part of the answer has to do with impending stock market valuation, speculative investment patterns and the difficulty of financing research with a 30-year output horizon, i.e., the optimistic predictions are partly about corporate boosterism. Part of the answer, however, has to do with an underlying difference in the final exam approach (which suggests 2042 is reasonable) versus power-scale approach (which suggests 2032 is reasonable). *Median categories of the proposed power scales ultimately correspond to occupationally relevant educational achievement levels*. If what you want is not an artificial simulation of the human intellect, but a machine learning interface that replaces intern-level and mid-level employees, then a graduate scale is likely to be congenial, and optimistic predictions are likely to seem realistic. In other words, while AGI is by no means all about instrumental rationality, the thing that you may be sold, or that might take your job, "AGI™," is most likely all about instrumental rationality.

An Alienation of the Human Intellect

Understanding what this means philosophically requires moving beyond technical description to the question of what kind of thing the human intellect actually is. From the Frankfurt School perspective, artificial intelligence is an externalization of the human intellect based on the objectification of human knowledge into a vast inventory of tendential connections (i.e., the model weights), which can presently be used either to simulate perception using pattern recognition on empirical data, or to simulate thinking and reasoning using categorial frameworks on textual inputs. When these are brought together, as is likely in the near future, the machine will model the cognitive operations that idealist philosophy attributed to transcendental subjectivity, applying categories to empirical data, in order to simulate the constitution of experience. As Jürgen Habermas reminds us, however, human knowledge is constituted historically and socially through embodied experience, in the species history of a natural animal with knowledge-constitutive interests in technical-cognitive control, social coordination through mutual recognition, and reflective self-knowledge under conditions of social freedom (Habermas, 2002: 301-317). The categories of thinking and reasoning that constitute objectivity are themselves constituted through the history of labor, not merely through the abstract process of their

deposition in some inventory. This is most evident in practical judgment, where the practical wisdom that decision-makers rely on arises through the learnt application of the art of judgment in social situations, in which context what is at stake is not user satisfaction, but mutual respect and evaluations of merit. What that means is that human knowledge—represented in the machine as a vast inventory of tendential connections—reflects an embodied social and historical process of discovery and innovation.

Human knowledge always returns to situated materiality in the lived experience that confirms or disconfirms those expectations which are formed through the sedimentation of knowledge in the process of socialization. This situated experience is temporalized—for perception and thinking happen irreversibly in the flow of the lived time of retention and protention. It is spatialized—for the human body has a being-there registered in the relevance of communication and the location of feelings. And it is qualitative—for perception and thinking are accompanied by "the feeling of what happens," the affective registration of the body in ideation. In the flow of experience, the human intellect holistically synthesizes perception, feeling, thinking and judgment—but most importantly, it does so against the background of the pre-reflective, tacit knowledge that forms the web of assumptions, meanings, values and tastes that constitute what Habermas calls the "lifeworld" (Habermas, 1987: 114-164). From the Frankfurt School perspective, as outlined by Habermas, a particular lifeworld as the horizon of lived experience cannot itself as a whole become the object of discussion—instead, regions of it are momentarily lit up, or thematized, in knowledge-constitutive debates and inquiries. The domains of human knowledge, expressed as vast ensembles of currently established claims to scientific truth, moral rightness and expressive authenticity, are embedded in lived experience with background knowledge, which contextualizes the knowledge claims as assumptions and know-how, without ever thematizing this directly.

It is against this background that the Frankfurt School restatement of the main objections to "Strong AI," the philosophical position on the conjectural technology of AGI that holds that it would have personhood status (and therefore moral rights), takes shape. According to Strong AI proponents, the goal of AI research is (or should be) to produce artificial persons, whereas according to Weak AI advocates, the aim is (or should be) to create simulations of persons (Bringsjord

and Govindarajulu, 2026). Notice that both of these positions assume the achievement not only of functional benchmarks for AGI, including the ability to pass the "Total Turing Test" (Harnad, 1991: 46), which includes hyper-realistic embodiment on the lines of the replicants in *Blade Runner*, but also credible evidence (or consistent simulations) of machine self-consciousness. According to John Searle, "Strong AI claims that thinking is merely the manipulation of formal symbols, and that is exactly what the computer does: manipulate formal symbols. Its view is often summarized by saying, 'The mind is to the brain as the program is to the hardware'" (Searle, 1990: 26). His original "Chinese Room" objection was that a rule-following, or algorithmically guided, manipulation process, happening in the absence of semantic understanding, could not count as thinking, even though it produced symbolic tokens (Searle, 1980: 418-419). But this was vulnerable to a powerful rejoinder. The rejoinder was that machine-learning based on stochastic processes inductively builds connections that model the brain's neural networks, in which thinking arises as an emergent property of the entire system (Bringsjord and Govindarajulu, 2026). The billions of synaptic connections do not think, it is the brain as a whole that thinks, using neurons and synapses, and a machine model of this process would think too. From the Frankfurt School perspective, however, this is not what thinking is, because human beings are not brains on sticks. In other words, the rejoinder misses not the architecture, but its grounding. Human beings are social animals, whose thinking happens by supplementing pre-reflective, lifeworld knowledge with explicit knowledge claims learnt in socialization processes—something which the machine simply cannot do.

The Frankfurt School position is adjacent to the hermeneutic objection of Hubert and Stuart Dreyfus, which was that thinking is not just a process of the symbolic modelling of the world, but grows from an inductive process of engagement with things, based on Being-there (Dreyfus and Dreyfus, 1987). Again, as stated, this position assumed that machine intelligence would involve programmed deductions, which led Hubert Dreyfus to significantly backtrack when confronted with "Heideggerian AI" based on deep learning (Dreyfus and Dreyfus, 1991, Dreyfus, 2007). But the original intuition, that the human intellect is something embodied and situated, something finite and natural, something that fuses perception, feeling, judgment and thinking in the flow of lived experience, remains entirely valid. The spectrum of meanings attached to

symbolic tokens is not simply tendential correspondences between commonly connected items in a sequence of words (i.e., a syntax), but felt expressions of personal relationships to contextually relevant worldly referents, uttered in the process of reaching understanding in order to coordinate action (i.e., a semantics) (Habermas, 1984: 99-100, 111-115). Yann LeCun's counter-proposal to LLM-based GenAI is instructive in this respect. According to LeCun, LLMs must regard the world as discrete because language tokenizes experience (LeCun, 2022). But the world is not discrete and definite—it is continuous and messy; the simulation of thinking performed by current LLM-based forms of GenAI must therefore artificially reduce the world to a kind of wireframe diagram whose nodes contain information, but whose gaps are nothingness. LeCun's proposed solution is fascinating: ground the models on simulations of perception based on modelling continuous distributions in messy environments and then equip them with world models, continually revisable pseudo-referential spaces, and self-models, which model identity as an inner world. Would this, perhaps, result in Strong AI?

The argument of Selmer Bringsjord and Michael Zenzen clarifies why the Frankfurt School position is emphatically "No!". As Bringsjord and Zenzen have shown, because computation is reversible, occurring in the atemporal space of calculation, but cognition is irreversible, transpiring in the temporalized space of thinking, AI cannot properly be understood as having thoughts (Bringsjord and Zenzen, 1997: 285, 313-314). It therefore cannot credibly accompany its representations with that apperceptive "I think" that Kant identified as the condition for unified experience, and so it has no experience, only outputs. The Weak AI position is the only credible position: the conjectural technology of AGI plus apparent consciousness would be a *simulation* of thinking, not a form of (human or inhuman) thinking. However, it presently shows no sign of consistent apparent consciousness, or of anything resembling an alien consciousness. It should be noted that often, the arguments and counterarguments in this space concern whether a machine could *in principle* be a being in terms of cognitive dignity and ethical status, which is an argumentative terrain whose gravitational field is strongly tilted towards conjectural technologies that belong to science fiction. With that proviso stated, their refutation of Strong AI is important, because many of the Strong AI positions verge on kinds of obfuscation whose main discursive significance is that they distract from the human consequences of the irresponsible development of this technology. For instance, a variant of the

Strong AI position is Alien AI, which is the claim that the goal of AI research is (or should be) a kind of conscious thinking which, although it might not pass a Total Turing Test (TTT), should have rights and responsibilities commensurate with its sapient status (Metzinger, 2021, Smith, 2025). Other sorts of Strong AI restatement depend on the claim that science has not solved the relation between neural substrate and conscious thinking, combined with the (contentious) assertion that functional performances express cognitive operations (Butlin, Long et al., 2023: 4). Such a machine has been seriously described in the philosophical literature as resembling what David Chalmers calls a zombie—a being whose outputs are indistinguishable from those of a conscious agent, but which entirely lacks interiority because it has no real emotions, inner experience or conscious awareness. I therefore take it that Bringsjord's proposal to create zombies—artificial person-like machines lacking inner experience or conscious awareness—beginning with zombanimals, which can be evaluated using test batteries such as TTT-rat, TTT-cat and TTT-dog (Bringsjord, Caporale et al., 2000: 397-398), in which the mechanical animal completely fools its real counterpart, is a kind of refutation using reductio of these sorts of Strong AI rejoinders. It sounds zany. But there is a serious point here. What the zombie proposal does is to attack the credibility of Strong and Alien AI arguments while simultaneously showing that even the Weak AI proposal (it is a simulation of conscious thinking) is way too strong.

Accordingly, the Weak AI position, that, whatever the functional performance of an AI, it is a machinic simulation of personhood not a human-equivalent artificial person, is not just a reasonable default in light of present objections to the Strong AI stance. Rather, it is a substantive position. From the Frankfurt School perspective, the conditions for genuine thinking (embodiment, temporality, and lifeworld embeddedness as ontological conditions; socialization, tool use, and interactive stakes as social conditions) are structurally unavailable to any machine simulation. Indeed, the degree to which the machine should even be grasped as performing a simulation of personhood needs to be critically delimited. Zombie is a name for it rich in connotations descending from the history of colonial slavery, which brings me now to the questions around the critique of the instrumental application of AI as a business machine designed to replace human labor.

Instrumental Reason and Mental Labor

The Frankfurt School position does not state that AI is nothing more than a baleful technology designed exclusively to replace humans with zombies, while reinforcing bias and wreaking environmental havoc. It states that AI is a severely limited objectification of the human intellect based on an externalization of human knowledge into what has rightly been called "stochastic parrots," which simulate perception and cognition without true interiority. There are nonetheless forms of AI, such as non-LLM foundation models that assist with predicting protein folding, designing therapeutic molecules, identifying radiological diagnoses, performing histopathology image analysis, classifying new enzymes, identifying cell types from single-cell analyses, weather forecasting, deforestation monitoring, … (this indicative list is very, very far from exhaustive) which are potentially beneficial to humanity. There are also LLM-based foundation models that perform potentially beneficial tasks: experimental design assistance, statistical code generation, automating accessibility affordances (such as automatic captioning), speech recognition and text prediction, OCR for historical texts, personalized tutoring and remedial essay-structuring advice, adaptive practice questions, conducting literature summaries … (this indicative list is also very, very far from exhaustive). With both LLM-based and non-LLM-based foundation models (Bommasani, Liang et al., 2022), then, there is a wide spectrum of applications where the technology is potentially valuable when appropriately developed and validated, and when responsibly powered and governed. It is therefore necessary to separate foundation models as a general purpose technology, representing a possible machine-learning technological revolution, from their controversial application through automation and surveillance, where "harmful" becomes a meaningful category, and from chatbots and generative media, where "pollution" has real traction. Most fundamentally, it is necessary to distinguish the beneficial potential of many applications of the technology from the social relations governing its implementation at the site of mental labor. AI becomes not just an objectification of the human intellect, but something opposed to the human being, when it is instrumentally deployed in the service of corporate profits and nationalist rivalries, as a means for the replacement of human labor and the destruction of human lives. Under these—currently prevailing—conditions, AI is an alienation of the human intellect.

The Deutscher Prize-winning social history of AI by Marxist historian Matteo Pasquinelli, entitled *The Eye of the Master* (2024), brings this critique into focus. The original "eye of the master" is the surveillance of the factory floor by the capitalist owner and their chief engineer, during the moment of the machine-action technological revolution and the labor agitation around the Machine Question in the 19th century, which abstracts the human movements of detail tasks into machine action. In the "Fragment on Machines" in the *Grundrisse*, Marx writes that "when, by the division of labor, each particular operation has been simplified to the use of a single instrument, the linking up of all these instruments, set in motion by a single engine, constitutes—a machine" (Pasquinelli, 2024: 105, Marx cited). The new "eye of the master" is the "automation of the psychometrics of labor and social behaviours—rather than the quest to solve the enigma of intelligence [it is the] automation of the statistical metrics which were originally introduced to quantify cognitive, social, and work-related abilities" (Pasquinelli, 2024: 244-245). In short, the machine-learning technological revolution performs, using stochastic processes that model complex behaviors, an extractive function in relation to mental labor analogous to that performed using machine actions that model simple movements in relation to manual labor. In a way that resonates with the Frankfurt School position on the emergence of knowledge from the lifeworld, he writes that "the act of image recognition or pattern classification is a specific kind of mental labor: it is a profoundly social act that mobilizes, at the same time, tacit and explicit know-how, scientific and traditional taxonomies, and vernacular and technical grammars" (Pasquinelli, 2024: 234). Pasquinelli sums this up by remarking that "for the way in which they have been built on large repositories of cultural heritage, collective knowledge, and social data, large foundation models are the closest approximation of the mechanization of the general intellect which was envisioned in the industrial age" (Pasquinelli, 2024: 248). His trenchant conclusion denounces "the operative principle of AI, [which] in fact, is not just labor automation but also the imposition of social hierarchies of manual and mental labor through automation," including the reinforcement of bias and the perpetuation of inequality as "an intrinsic discriminatory feature of automation in a capitalist context" (Pasquinelli, 2024: 21).

In contesting alienation through automation and the pervasive reduction of the human intellect to instrumental calculations, however, it is crucially important not to replicate instrumental reason by reducing human thinking to

mental labor. Even though for Marx, the concept of labor has been lent anthropologically general scope, as creative praxis, as Marxists such as Gyorgy Lukács recognized, labor is irreducibly not just teleological action but a teleological manipulation of natural material (Lukács, 1980), i.e., instrumental action. Habermas's critique of the reduction of action to just instrumental manipulation and thinking to just means-ends calculations of efficiency and effectiveness (Habermas, 1975) makes it possible to point out that the human intellect also engages in cognitive curiosity, ethical reasoning, the art of judgment, the expression of authentic feelings, and artistic play. The reduction of an externalization of human knowledge to business tasks and military strategies is something that should be deplored, not accepted as intrinsic to the technology or a representation of what thinking, in its richness, actually is. The real point is not that AI is exclusively an automation technology, but that its externalization of human knowledge is based on quantification and instrumentalization in ways that make these applications the most likely implementation pathway. Here, Pasquinelli is correct to claim that this has been prepared by the division of labor that resulted from the ICT Revolution, which integrated and digitized informational flows while generating big data, making white-collar workplace surveillance possible while opening some forms of mental labor to statistical analysis. To this, it should be added that—judging by past waves of automation (Braverman, 1974)—a likely outcome of the implementation of AI as a productivity technology will be the homogenization and routinization of mental labor itself.

The Microeconomics of AI Diffusion

The development and implementation of AI is happening under conditions shaped by two imperatives (Deranty and Corbin, 2024: 676-677). (1) The capitalist imperative means that AI development is competitively oriented and that AI implementation is directed by the profit motive. The capitalist imperative dictates that an uneven pattern of AI implementation, similar to previous waves of automation, is likely to happen. Decisions between the augmentation or the displacement of workers in the production process will be considered from the perspective of profitability—and not, for instance, of the social cost or the human consequences. (2) The nationalist imperative means that AI is regarded from the perspective of national advantages and military rivalries. The nationalist

imperative militates against AI safety because it locks AI research into a race for asymmetrical warfare potentials and competitive formulations of marginal advantage. This means that frontier models are unlikely to be completely free from alignment problems. Meanwhile, media discussion of the potential impact of AI on the workplace typically concentrates on spectacular scenarios of technological unemployment (e.g., "AI will replace 40% of jobs by 2030" (Altman, 2025)), although there is some recent recognition of the potential for productivity augmentation. The likely impact of AI implementation in the near future can be glimpsed from the study behind the headlines, which measured AI implementation relative to potential implementation in the 10 most highly impacted occupational categories. According to Anthropic's Maxim Massenkoff and Peter McCrory, "[i]f all workers within the top 10% of coverage were laid off, it would increase unemployment within the top quartile group from 3% to 43%, and it would increase aggregate unemployment from 4% to 13%" (Massenkoff and McCrory, 2026: 12). This is very serious. "The joint effect of a large influx of displaced workers at the seemingly unique inflection point posed by AI, together with their longer life expectancy, may thus create a formidable challenge that even the most advanced welfare state will be hard pressed to cope with," writes Manuel Trajtenberg. "Put differently, we cannot afford to have many more, and longer-lived, unemployed or underemployed people [but] this is what is at stake with the advent of AI" (Trajtenberg, 2019: 178).

Economic change caused by technological innovation is normally a protracted process characterized by belated diffusion and delayed impacts (Smil, 2005, Smil, 2006). The disruptive effects of technological innovation typically start very slowly but then last more than a decade for highly-exposed sectors, with ripples propagating through the rest of the economy for another decade or more, including, typically, collapses of speculative bubbles (e.g., the dot-com bubble) and periods of high unemployment. The "productivity paradox" of the Information and Computer Technology (ICT) Revolution in the 1980s led Nobel Prize winner Robert Solow to famously exclaim that "you can see the computer age everywhere but in the productivity statistics" (Acemoglu, Autor et al., 2014: 394 Solow cited). The "ICT productivity paradox" sprang from the fact that despite significant investment in ICT by most major corporations during the 1980s, productivity statistics and profit margins remained stagnant until the 1990s, when the propagation of computers throughout big firms was joined by

economy-wide transformations. A new "productivity paradox" is emerging with AI. This has both microeconomic and macroeconomic aspects. The microeconomic aspect is that AI often boosts individual productivity, but has yet to improve economic productivity statistics or to unequivocally impact on employment data (Gans, 2026). The macroeconomic aspect is that financial investment in AI firms is not proportional to profit improvements following early AI diffusion, leading to concerns about a speculative bubble and uncertainty about whether or not a paradigm shift is happening (Acemoglu, 2024). This AI productivity paradox has invited light-minded dismissals of the potential disruption that AI might cause, based on the idea that, because it has failed to meet excessive expectations, it will not be widely adopted or might evaporate following a speculative collapse. Just as with the ICT Revolution, the AI Revolution may be expected to last decades, involving significant labor process redesign, skill migration to new jobs and major reskilling across the workforce, organizational restructuring consistent with platform capitalism, enterprise system re-engineering and the development of AI-guided business models. It may also be expected to involve both technological unemployment and AI augmentation, in a disruptive period of economic and social adjustment that I speculate might perhaps (because the nature of AI militates in the direction of the routinization and standardization of labor processes to improve their predictability) involve a shift from Post-Fordism to Neo-Fordism. The AI productivity paradox is unlikely to remain paradoxical for long.

The report by Massenkoff and McCrory, "Labour Market Impacts of AI" (2026) is not corporate advertising for the further deployment of Anthropic's product, Claude, but rather an analysis of why there is a difference between how Claude is used and what is happening in the workplace, which they describe as "observed coverage". They explain the proposed measure of observed coverage in terms of how "we count tasks that are theoretically capable with an LLM as covered if they have seen sufficient work-related usage in Claude traffic," with "theoretical capability" defined by a "measure of whether it is theoretically possible for an LLM to make a task at least twice as fast" (Massenkoff and McCrory, 2026: 5-6). There might be a variety of reasons for a big gap between theoretical capability and observed coverage, including conceptual and practical uptake by employees, but also, and probably more significantly, "legal constraints, specific software requirements, human verification steps, or other

hurdles" (Massenkoff and McCrory, 2026: 5). The relationship between this gap and the totality of tasks performed by occupational category is measured in the study in terms of occupational "exposure". It should be noted that "*exposure is typically defined at the task level*: AI can grade homework but not manage a classroom, for example, so teachers are considered less exposed than workers whose entire job can be performed remotely" (Massenkoff and McCrory, 2026: 3 emphasis added). There is a small and well-known number of groups of occupational categories that are widely regarded as "canaries" for the process of the replacement of mental labor by machine-learning processes, such as data entry and software coding, financial and accounting services, paralegal assistants, and some categories of customer service representatives. A 2025 study showed a 16% decline in hires of early career workers (aged 22-25) in these occupational categories as something that could be confidently correlated with AI exposure (Brynjolfsson, Chandar et al., 2025). This finding lends support to the position of Massenkoff and McCrory that although AI effects are difficult to isolate, good reason exists to suppose that its impact in high exposure categories will be strongly felt in the near future.

The central reason why the diffusion and implementation of Generative AI is unlikely to result solely in technological unemployment is that the technology splits prediction from decision in the process of judgment (Agrawal, Gans et al., 2019, Agrawal, Gans et al., 2023, Agrawal, Gans et al., 2025). Joshua Gans and co-thinkers argue that AI creates a new skill continuum, between low-stakes prediction-intensive mental labor (e.g., record keeping) and high-stakes decision-intensive mental labor (e.g., lawyer, professor). The implementation of AI is likely to cause technological unemployment amongst low-stakes prediction-intensive mental laborers, but expert augmentation amongst high-stakes decision-intensive mental laborers.

In workplace settings where the vast majority of task elements involve low-stakes prediction-intensive performances, and labor costs are the most significant factor affecting profit margins, three things are observed (Agrawal, Gans et al., 2025):

1. AI augments the capacity of low-skill workers, raising their productivity towards the average;
2. AI does not significantly improve the productivity of high-skill workers;

3.Entry-level workers (skill unknown) are not cost effective.

Here, managements might consider (1) augmenting mid-ranked, mid-pay, mid-skill laborers with AI while holding their wages relatively static (2) displacing high-ranked, high-pay, high-skill workers to save costs and (3) outsourcing surplus tasks via a platform to casualized low-pay, unknown skill workers (formerly entry-level aspirants and displaced former workers). These are the considerations that drive debate on technological unemployment.

Notice that I assume that the displacement of both high-skill and low-skill workers here will drive further platformization and casualization—this assumption is *not* articulated by Gans and co-thinkers, and probably not shared by them.

By contrast, where the vast majority of task elements involve high-stakes, decision-intensive performances, especially in relation to complex task ensembles (e.g., scientific research, legal practice, medical practice), and labor costs are the most significant factor affecting profit margins, it is observed that (Agrawal, Gans et al., 2025):

1.AI augments the capacity of high-skill workers, raising their productivity above the average;

2.AI does not significantly improve the productivity of low-skill workers;

3.Entry-level workers are not cost effective, unless already high-skill.

Here, management might consider (1) augmenting high-ranked, high-pay, high-skill laborers with AI while holding their wages relatively static (2) demoting high-pay low-skill laborers to paraprofessional status while increasing the scope of paraprofessional responsibilities and (3) demanding that entry-level workers exhibit high skill as a condition of employment. These are the considerations that drive the arguments about expert augmentation.

The differential impacts of AI implementation amongst low-stakes prediction-intensive versus high-stakes judgment-intensive occupational categories can be expected to produce a sharp polarization of the white-collar workforce. But the prediction-intensive versus judgment-intensive continuum is not restricted to occupational categories—as Massenkoff and McCrory note, "exposure is typically defined at the task level". What will the effects be when this continuum is introduced *within* every occupational category? Aggregation into

the center, with devolution of decision-making from professional onto paraprofessional strata resulting in improved prestige and earnings, or proletarianization of the bottom and erosion of the middle, with the automation of routine tasks differentially impacting occupational strata?

The Structural Transformation of the Middle Class

The liberal economic papers of labor market economist David Autor helpfully focus this question around the sort of "rebuilding the middle class" scenario (not prediction) that the Frankfurt School critique needs to engage seriously, not just dismiss. For Autor, under political conditions of pro-social regulation, the diffusion of AI "can be applied to enable a larger set of workers possessing complementary knowledge to perform some of the higher-stakes decision-making tasks that are currently arrogated to elite experts" (Autor, 2024: 2). Autor's paradigmatic example is the creation of the paraprofessional role of Nurse Practitioner (NP), a category of Registered Nurse (RN) whose additional Master in Nursing degree qualifies them to perform tasks—"administer and interpret diagnostic tests, assess and diagnose patients, prescribe medications"—formerly restricted to doctors (Autor, 2024: 8-9). In the study, Autor acknowledges that the creation of NPs in the 1960s was facilitated by information technology, not GenAI, and involved government surmounting the political resistance of the AMA. Nonetheless, he points out that the category tripled in job numbers in the period 2011-2022, partly as a result of ICT integration, and now commands a well-above average salary in the range of US $125,900 P.A., making these employees "elite decision-makers". "Their work combines procedural expertise with expert judgment so they can confront one-off patient cases where the stakes for judicious decision-making are extraordinarily high," providing a "large-scale case where high-stakes professional tasks—diagnosing, treating and prescribing—have been reallocated (or co-assigned) from the most elite professional workers (MDs) to another set of professionals (NPs) with somewhat less elite (though still substantial) formal expertise and training" (Autor, 2024: 9). For Autor, this provides a template for how AI plus re-education can augment paraprofessional strata through the devolution of some tasks from elite professionals onto lower layers, lifting up the middle strata through AI augmentation, rather than casting them down through technological unemployment. (The same sort of argument, albeit with a more controversial case study, might have been made in relation to

graduate university tutors engaged as non-research Education Focused Academics.) According to Autor:

> By providing decision support in the form of real-time guidance and guardrails, AI could enable a larger set of workers possessing complementary knowledge to perform some of the higher-stakes decision-making tasks currently arrogated to elite experts like doctors, lawyers, coders and educators. This would improve the quality of jobs for workers without college degrees, moderate earnings inequality, and—akin to what the Industrial Revolution did for consumer goods—lower the cost of key services such as healthcare, education and legal expertise (Autor, 2024: 8).

It is worth engaging closely with the texture of Autor's "complementarity argument" to understand why the Frankfurt School critique suggests that, although Autor's scenario is desirable, the opposite—intense polarization—is equally (or perhaps more) likely. The key point in Autor's argument is that, citing Agrawal, Gans and Goldfarb on the high-stakes judgment-intensive versus low-stakes prediction-intensive continuum, AI is an "inversion technology" (Autor, 2024: 8). What he means by that, however, is something quite different to the microeconomics of AI position, for according to Autor, an "inversion technology" is a "complementary" one that inverts the normal "substitution" pattern of automation (which de-skills the majority and replaces a minority) (Autor, 2024: 12). AI is complementary to workers who have solid foundation skills but lack elite educational credentials because "in a cosmic irony, AI is not trustworthy with facts … [and] does not respect rules" but is "remarkably effective at acquiring tacit knowledge" (Autor, 2024: 7). It is "like a jazz musician, riffing on existing melodies, taking improvisational solos and humming new tunes," in short, acting like a human expert to weave formal knowledge with acquired experience to make judgments about singular cases—*but*, unlike the human jazz musician, it lacks "classical musical training," that is, it is hopeless with foundational knowledge and basic training (Autor, 2024: 7). Consequently, AI provides expert-level top-ups to paraprofessional human workers whose "classical" skills provide the solid foundation that makes AI implementation a responsible form of augmentation not a destructive form of substitutive automation.

That extraordinarily optimistic argument depends strongly on its definition of the relation between expert judgment and tacit knowledge, by which

latter Autor means Michael Polanyi's definition of implicit "knowing-that-and-how-to," operationalized for instance in riding a bicycle. Polanyi's book proposes paradigms centered on pattern recognition (e.g., noticing a mood from a facial expression, recognising a symptom in a patient's comportment) and skilled performance (e.g., an athletic performance, assisted walking with a mobility cane) (Polanyi, 1966: 1-26). These instances illuminate the problem with Autor's conception of expert judgment, which conflates pattern recognition and syntactical expertise (which are prediction-intensive activities that AI plausibly simulates) with deliberative decision-making, under high-stakes conditions of significant uncertainty with major consequences for an incorrect output. In terms of Autor's jazz musician analogy, what AI simulates is the sort of artist who can perform regulated improvisations, as opposed to the sort of artist who can make original music or freely adapt their melodies to the situation. By contrast, the framework of Gans (and co-thinkers) explicitly separates these two aspects of decision-making into prediction-intensive and judgment-intensive, while specifying that what "high-stakes" and "low-stakes" mean relates to output valuation—knowing what matters, what an error costs and what the prediction entails. By claiming that AI complements foundation skills with expert judgment based on tacit knowledge, then, Autor misses the activity of valuation which is the core (from the microeconomic perspective) of judgment. In the case of the Nurse Practitioners (NPs), for instance, expert judgment in mid-stakes judgment-intensive decision-making is exactly what the Masters in Nursing is for, relative to the credential required for a Registered Nurse (RN), and precisely what AI does *not* do particularly well. The suggestion that AI might "top up" an RN to NP level is not particularly persuasive.

The microeconomic framework predicts differential impacts in the paraprofessional realm—something which the examples of NP (Autor) and radiologist (Agrawal and co-thinkers) illustrate. In the case of the NP, what is devolved is a set of decision-making tasks which involve some guidance from prediction-intensive routines (i.e., differential diagnoses), but where these merely supplement procedural skill, embodied care work, escalation decisions, and legally structured accountability. The automation of differential diagnoses can be predicted to augment the role of an NP. Something entirely different happens with the radiologist, whose task bundle involves a majority of prediction-intensive routines (also involving differential diagnoses). Here, what Agrawal

and co-thinkers found was that typically human radiologists overruled correct machine inferences, because their training had not prepared them to understand that human intuition is sometimes inferior to machine forms of pattern recognition (Agrawal, Gans et al., 2019). The calibrated deference that this development implies is unlikely to lead to job enhancement and although the predicted extinction of radiology as a profession has not happened (and likely will not), the degradation of the status and salary of radiologists is a real possibility. In short, the microeconomic framework tends to predict polarization between and within professions, as prediction-intensive tasks succumb to automation (with associated degradation of labor) and judgment-intensive tasks become more valuable. The occupations associated with those tasks that have been automated typically do not cease to exist, but instead become more widespread (initially) because the task performance has become cheaper (and therefore more available). But that is unlikely to be a good thing for the human employees.

A much darker scenario emerges when the relation between "silicon empires" (Srnicek, 2025) and "platform capitalism" (Srnicek, 2017) is considered. Sarah Burkhardt and Bernhard Rieder argue persuasively that "foundation models are platform models," because a foundation model is the basis for a wide variety of (business) applications that can then be integrated, through "deep user interactions, reordered cost structures, and centralization and lock-in" into a platform-based gig economy (Burkhardt and Rieder, 2024: 8-12). The highly exposed occupational categories, and the residues of other occupational categories once task devolution has happened, can easily end up casualized or contracted through platforms. Because foundation models not only make platformization possible, but also involve control over cloud computing (Narayan, 2022), plus vertical integration, complementary innovation and the power of abstraction (Luitse, 2024), it is possible that a new economic paradigm is emerging around what are basically technological rents (Rikap, 2021).

Discussion of the possibility of neo-Fordism takes me too far from the scope of the present article, but I will close by stating that the internal polarization of mental labor as a result of AI implementation seems most likely set to resemble the proletarianization of office work in the C20th. It would take a significant political struggle for it to produce a new era of comfort in the middle strata. Think about the business model of the manufacturers of foundation models, which is

corporate subscriptions for labor-saving technology. Only this revenue can provide a return on truly vast speculative investment: not private subscriptions, not embedded advertising, but the trillion dollars annually slashed from the budgets of the Fortune-500 companies, through AI substitution and robotic automation across 90% of occupations (Morgan Stanley Research, 2025).

What this points to, alongside platformization into the gig economy and workplace speed-up, is the potential for a structural transformation of the middle class, including casualization and cheapening of highly exposed occupational categories. This may be accompanied by the collapse of some credential types and the erosion of those social allegiances that have historically linked the lower middle class to status quo political parties. Social dislocation and political convulsion on the order of the 1930s is far from excluded as a conjunctural possibility in the timeframe (2027-2037) predicted by the AI CEOs and now echoed by the World Bank and the ILO.

Coda: Catastrophic Risks

Apparently, Anthropic CEO Dario Amodei recently increased his "p(doom) number" to about a 25% probability of catastrophe (Morrone, 2025)—a figure which is broadly in line with the concerns of other CEOs of frontier AI corporations, such as Elon Musk and Sam Altman (Kokotajlo, Alexander et al., 2025). Demis Hassabis, who thinks that guessing probabilities of catastrophe scenarios is absurd, nonetheless also warns that "it's definitely non-zero and it's probably non-negligible," and that dismissing existential risk is "ridiculous" (Hamill, 2025). It is difficult to overstate how important it is that the designers of this technology themselves not only constantly warn that an AI arms race in which safety research is significantly eroded is recklessly dangerous, but that the supposedly benign effects of their instrument—as a labor-saving business machine—could involve massive social dislocation. I believe that in the history of technology, only the makers of the atomic bomb have been that anxious about the potentially destructive effects of their own invention; the idea of just ignoring their warning or dismissing it as hype can only be described as unwise.

The basket of problems alluded to by Amodei's p(doom) estimate is significantly different from the "AI Doomer" scenarios that he distances himself from in his essay on the potential "failure modes" of "adolescent" AGI (Amodei, 2026). For prophets of doom such as Eliezer Yudkowsky—"if anybody builds it,

everybody dies" (Yudkowsky and Soares, 2025)–as well as for more reasonable, if still hyperbolic, critics such as Geoffrey Hinton, Nick Bostrom (Bostrom, 2002, Bostrom, 2016) and Daniel Kokotajlo (and co-thinkers) (Kokotajlo, Alexander et al., 2025), the problem is misaligned superintelligence. Such scenarios all involve a rapid transition to AGI, followed by accelerated (indeed, runaway) development to ASI, as a result of self-recursive improvement happening at billions of times the speed of human-gated improvement processes. (In a typical scenario, for instance, this basically happens over the weekend, when some junior researcher fails to sit anxiously by the self-destruct button while the machine supervises its own acceleration.) Just to be clear: if that were possible, ASI would be extremely dangerous, and misaligned ASI would indeed pose an existential risk. However, in light of the real-world performance of LLMs against demanding AGI benchmarks, stealth-ASI scenarios are presently just anxiety-fantasies, which distract from the actual problems associated with this technology. Now to AGI. Perhaps the CEOs of the AI corporations, who have access to the most advanced frontier models, know more than the scientists who administer tests such as ARC-AGI-Eval and Humanity's Last Exam. Perhaps AGI really is just around the corner (2027-2032), because self-recursion makes it possible to evade the energy-scaling divergence, or because what is being sent out for testing is already several generations behind what is locked up in the lab. Perhaps. What is certain is that for Amodei, as for the other CEOs who have expressed alarm about AI, AGI is a very-near-future technology that should be intensely safeguarded and approached carefully, because although its effects are difficult to calculate, its disruptive consequences are easy to predict. Amodei's p(doom) basket is not an ASI breakout scenario, but a bundle of concerns around military applications, technological unemployment, the absence of a science of AI alignment and the possibility of malign misuse (Amodei, 2026).

The simple truth is that it is not necessary to agree with Amodei's AGI timeline to accept the wisdom of imposing strong safety and social regulations on AI development. These might include an international convention preventing ASI research, verifiable frameworks for AI-research transparency, ethical guidelines on the safeguards that need to be built into AGI, a massive research slowdown to discover the science of AI-control, environmental safeguards on datacenter provision, and a treaty banning the delegation to AI of nuclear decisions. Some of the AI CEOs advocate this already—others actively undermine it daily. But such

agreements and conventions only touch the surface. It is easy to say that AI research should be not just nationalized, but something delegated upwards to the international community, not in the hands of corporations and governments. I fear, though, that these are just words, when what is needed is an active struggle that aims, not to reassure individuals about their political virtue, but to actually win concrete demands. The most difficult yet pressing issue is the political regulation of AI implementation in the workplace. Here, something like "complementarity laws" preventing substitutive automation might play a role in ensuring pro-social outcomes. Equally urgent is a treaty banning autonomous weapons systems, including autonomous drones, implementing something like an Asimov Principle. Really, even if the Three Laws don't actually work, and what is needed is something akin to a "constitution" for each AI agent, this should be built into every AI as a matter of course. But the foundation models themselves need scrutiny, not just in terms of the intellectual property violations that constitute their training data, but also in terms of a rigorous effort to eliminate bias through corrective re-training or another post-training process. Finally, proposals for slowdown until a science of control is developed should be taken seriously, along with practical suggestions for transparency agreements and research safeguards (Kokotajlo, Larsen et al., 2026).

It needs no superintelligence, escaped from the lab, to have very real plans to incinerate the planet and exterminate humanity—our own generals have just such plans in the event of nuclear escalation, and they, unlike a conjectural silicon antichrist, have the means to enact them, including present-day AI tools that "improve" the probability of megadeaths. Nor does it require the arrival of AGI to cause massive social dislocation because of technological unemployment, educational dysfunction, democratic erosion, political radicalization and nationalist agitation—the very corporations whose CEOs talk AGI regulation are presently releasing sub-AGI models into the workplace with historically disruptive effects. Ignoring or dismissing catastrophe scenarios is unwise. But fixating on them when the mundane reality of AI diffusion calls for urgent action is just a distraction.

References

Acemoglu, Daron (2024) "The Simple Macroeconomics of AI". MIT Faculty of Economics, MIT: 1-50.

Acemoglu, Daron, *et al.* (2014) "Return of the Solow Paradox?" *American Economic Review* 104(5): 394-399.

Agrawal, Ajay, *et al.* (2019) Prediction, Judgment, and Complexity: A Theory of Decision-Making and Artificial Intelligence. *The Economics of Artificial Intelligence: An Agenda (National Bureau of Economic Research Conference Report)*. A. Agrawal, J. Gans and A. Goldfarb. Chicago, University of Chicago Press: 89-114.

Agrawal, Ajay, *et al.* (2023) *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Boston, Harvard Business Review Press.

Agrawal, Ajay, *et al.* (2025) *Power and Prediction: The Disruptive Economics of Artificial Intelligence*. Boston, Harvard Business Review Press.

Altman, Sam (2025) "Sam Altman explains why AI will replace 40% of your work". *Business Insider*. E. S. Domena. Berlin; New York, Business Insider (Springer).

Amodei, Dario. (2024) "Machines of Loving Grace: How AI Could Transform the World for the Better. "Retrieved 6 Jan 2026 from https://www.darioamodei.com/essay/machines-of-loving-grace.

Amodei, Dario. (2026) "The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI." Retrieved 7 July 2026 from https://darioamodei.com/essay/the-adolescence-of-technology.

ARC-AGI-Prize, The. (2026) "Claude Opus 5 (24 July 2026)." Retrieved 30 July 2026 from https://arcprize.org/results/anthropic-claude-opus-5.

Autor, David (2024) "Applying AI to Rebuild Middle Class Jobs". Cambridge, MA, National Bureau of Economic Research: 1-22.

Bender, Emily M., *et al.* (2021) "On the Dangers of Stochastic Parrots: Can Large Language Models Be Too Big?". *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM Digital Library: 610-623.

Benjamin, Ruha (2019) *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge, Polity Press.

Bergmann, Dave and Cole Stryker. (2026) "What Is Artificial General Intelligence (AGI)?" Retrieved 20 Jan 2026 from https://www.ibm.com/think/topics/artificial-general-intelligence.

Bommasani, Rishi, *et al.* (2022) "On the Opportunities and Risks of Foundation Models". Stanford, Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University: 1-214.

Bostrom, Nick (2002) "Existential Risks." *Journal of Evolution and Technology* 9(1): 1-31.

Bostrom, Nick (2016) *Superintelligence: Paths, Dangers, Strategies (Reprint Edition)*. Oxford, Oxford University Press.

Braverman, Harry (1974) *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York, Monthly Review Press.

Bringsjord, Selmer, *et al.* (2000) "Animals, Zombanimals, and the Total Turing Test: The Essence of Artificial Intelligence." *Journal of Logic, Language, and Information* 9(4): 397–418.

Bringsjord, Selmer and Naveen Sundar Govindarajulu (2026) "Artificial Intelligence". *The Stanford Encyclopedia of Philosophy (Summer 2026 Edition)*. E. N. Zalta and U. Nodelman: online.

Bringsjord, Selmer and Michael Zenzen (1997) "Cognition Is Not Computation: The Argument from Irreversibility." *Synthese* 113(2): 285-320.

Brynjolfsson, Erik, *et al.* (2025) "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence". Stanford, CA, Stanford Digital Economy Lab: online.

Burkhardt, Sarah and Bernhard Rieder (2024) "Foundation Models are Platform Models: Prompting and the Political Economy of AI." *Big Data and Society*(April-June): 1-15.

Butlin, Patrick, *et al.* (2023) "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness (Pre-press Version)." *ArXiv*: 1-88.

Chollet, François (2019) "On the Measure of Intelligence." *arXiv*: 1-64.

Chollet, François, *et al.* (2026) "ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence". online, ARC Prize: 1-23.

Chollet, François, *et al.* (2025) "ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems (Version-2)." *ArXiv*: 1-13.

Chun, Wendy (2024) *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA, MIT Press.

Coeckelbergh, Mark (2022) *The Political Philosophy of AI*. Cambridge, UK, Polity Press.

Deranty, Jean-Philippe and Thomas Corbin (2024) "Artificial Intelligence and Work: A Critical Review of Recent Research from the Social Sciences." *AI & Society* 39: 675-691.

Dilmegani, Cem and Sıla Ermut. (2026) "AGI/Singularity: 9,800 Predictions Analyzed." Retrieved 7 JUL 2026, from https://aimultiple.com/artificial-general-intelligence-singularity-timing.

Dreyfus, Hubert (2007) "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian." *Artificial Intelligence* 171(18): 1137-1160.

Dreyfus, Hubert and Stuart Dreyfus (1987) *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York, Free Press.

Dreyfus, Hubert and Stuart Dreyfus (1991) Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at the Branchpoint. *Understanding the Artificial: On the Future Shape of Artificial Intelligence*. M. Negrotti. Berlin, Springer Nature: 33-54.

Fuchs, Christian (2022) *Digital Humanism: A Philosophy for 21st Century Digital Society (Kindle Edition)*. Bingley, UK, Emerald Publishing.

Gans, Joshua (2026) *The Microeconomics of Artificial Intelligence*. Cambridge, MA, MIT Press.

Goertzel, Ben and Cassio Pennachin (2007) Contemporary Approaches to Artificial General Intelligence. *Artificial General Intelligence*. B. Goertzel and C. Pennachin. Berlin, Springer: 1-30.

Gray, Mary and Siddharth Suri (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York, Houghton Mifflin Harcourt.

Habermas, Jürgen (1975) "Towards a Reconstruction of Historical Materialism." *Theory and Society* 2(3): 287-300.

Habermas, Jürgen (1984) *The Theory of Communicative Action: Reason and the Rationalisation of Society*. Boston, Beacon Press.

Habermas, Jürgen (1987) *The Theory of Communicative Action: System and Lifeworld*. Boston, Beacon Press.

Habermas, Jürgen (2002) *Knowledge and Human Interests*. Boston, Beacon.

Hamill, Jasper (2025) "Google Deepmind Boss Demis Hassabis Reveals His P(Doom)". *Machine*. London, Machine Magazine: online.

Hao, Karen (2026) *Empire of AI: Inside the Reckless Race for Total Domination*. London; New York; Sydney, Penguin.

Harnad, Stevan (1991) "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem." *Minds and Machines* 1(1): 43–54.

Johnson, Matthew (2025) "AGI Has Quietly Become Central to Beijing's AI Strategy." *China Brief (Jamestown Foundation)* 25(18): 14-20.

Kokotajlo, Daniel, *et al.* (2025) "AI 2027". Sacramento, CA, AI Futures Project.

Kokotajlo, Daniel, *et al.* (2026) "AI 2040: Plan A". online, AI Futures Project: 1-90.

LeCun, Yann (2022) "A Path Towards Autonomous Machine Intelligence." *Open Review dot Net*: 1-62.

Levesque, Henry (2014) "On Our Best Behaviour." *Artificial Intelligence* 212: 27–35.

Luitse, Dieuwertje (2024) "Platform Power in AI." *Internet Policy Review* 13(2): 1-12.

Lukács, György (1980) *The Ontology of Social Being: Labour*. London, Merlin.

Maslej, Nestor , *et al.* (2025) "The AI Index 2025 Annual Report". A. I. S. Committee. arXiv, Stanford University Institute for Human-Centered AI: 1-457.

Massenkoff, Maxim and Peter McCrory (2026) "Labor Market Impacts of AI: A New Measure and Early Evidence". San Francisco, CA, Anthropic: 1-17.

Metzinger, Thomas (2021) "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology." *Journal of Artificial Intelligence and Consciousness* 8(1): 43-66.

Morgan Stanley Research. (2025) "AI Could Affect 90% of Occupations." 17 SEP 2025. Retrieved 31 JUL 2026, from https://www.morganstanley.com/insights/articles/ai-workplace-outlook-2H-2025.

Morrone, Megan. (2025) "Amodei on AI: There's a 25% Chance Things Go Really Badly." 17 SEP 2025. Retrieved 31 JUL 2026, from https://www.axios.com/2025/09/17/anthropic-dario-amodei-p-doom-25-percent.

Narayan, Devika (2022) "Platform Capitalism and Cloud Infrastructure: Theorizing a Hyper-scalable Computing Regime." *Environment and Planning* 54(5): 1-15.

Noble, Safiya (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, New York University Press.

O'Neil, Cathy (2017) *Weapons of Math Destruction*. London; New York; Sydney, Penguin.

OpenAI. (2026) "OpenAI Charter." Retrieved 27 JAN 2026, 2026, from https://openai.com/charter/.

OpenAI (Achiam et al.), Development Team and Adrien Ecoffet (Corresponding Author) (2024). "GPT-4 Technical Report 2024v6." *arXiv*: 1-100.

Pasquinelli, Matteo (2024) *The Eye of the Master: A Social History of Artificial Intelligence*. London; New York, Verso.

Phan, Long, *et al.* (2025) "Humanity's Last Exam." *arXiv*: 1-29.

Polanyi, Michael (1966). *The Tacit Dimension*. New York, Doubleday & Co.

Rikap, Cecilia (2021) "From Global Value Chains to Corporate Production and Innovation Systems: Exploring the Rise of Intellectual Monopoly Capitalism." *Area Development and Policy* 7(2): 147-161.

Ringel Morris, Meredith , *et al.* (2025) "Levels of AGI for Operationalizing Progress on the Path to AGI v5." *arXiv*: 1-14.

Roser, Max. (2023, 7 FEB 2023) "AI Timelines." Retrieved 23 JAN 2026, 2026, from https://ourworldindata.org/ai-timelines.

Searle, John (1980) "Minds, Brains and Programs." *The Behavioural and Brain Sciences* 3: 417-457.

Searle, John (1990) "Is the Brain's Mind a Computer Program?" *Scientific American* 262(1): 25-31.

Smil, Vaclav (2005) *Creating the Twentieth Century: Technological Innovations of 1867-1914 and their Lasting Impact*. Oxford, Oxford University Press.

Smil, Vaclav (2006) *Transforming the Twentieth Century: Technological Innovations and their Consequences*. Oxford, Oxford University Press.

Smith, Kevin H. (2025) "Artificial Minds, Alien Experiences: Navigating the Uncharted Territory of AI Phenomenology." *University of Memphis Law Review* 55: 621-749.

Srnicek, Nick (2017) *Platform Capitalism*. Cambridge, Polity Press.

Srnicek, Nick (2025) *Silicon Empires: The Fight for the Future of AI*. London, Polity.

Tihanyi, Norbert, *et al.* (2024) "Dynamic Intelligence Assessment: Benchmarking LLMs on the Road to AGI with a Focus on Model Confidence." *ArXiv*: 1-9.

Trajtenberg, Manuel (2019) AI as the Next GPT: A Political- Economy Perspective. *The Economics of Artificial Intelligence: An Agenda (National Bureau of Economic Research Conference Report)*. A. Agrawal, J. Gans and A. Goldfarb. Chicago, University of Chicago Press: 175-186.

Turing, Alan (1950) "Computing Machinery and Intelligence,." *Mind* LIX: 433–460.

Yudkowsky, Eliezer and Nate Soares (2025) *If Anyone Builds It, Everyone Dies*. Boston, Little, Brown & Company.

Zhong, Wanjun , *et al.* (2023) "AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models v-2." *arXiv*: 1-22.

M–I–M′: AI and The Informatic Violence of Capital

Jonathan Beller[1]

Abstract: This essay develops a theory of "computational racial capitalism" through a reformulation of Marx's general formula for capital, M–C–M′, as M–I–M′: money–information–more money. It argues that information has become a paradigmatic means of accumulation, through which capital increasingly converts social, material, and biological differences into quantities that can be calculated, exchanged, and valued. Computation is therefore not merely a tool of contemporary capitalism but an increasingly constitutive form of capital itself. The essay develops the concepts of "economic media," "informatic labor," and "living information" to describe this transformation. As computation subsumes prior media and renders ever more domains of life as information, human activity becomes a continuous source of machine-readable differences and state changes through which capital expands.

This process is inseparable from the racialized devaluation of life. Surveillance, policing, border regimes, incarceration, forced displacement, and warfare become significant sites for the extraction of living information precisely where racial capitalism has historically discounted certain lives and populations through enslavement and colonization. The genocide in Gaza provides a particularly stark instance of the convergence of computation,

[1] **Jonathan Beller** is Professor of Humanities and Media Studies and co-founder of the Graduate Program in Media Studies at Pratt Institute. He is also Distinguished Visiting Professor of English and of Women's, Gender, and Sexuality Studies at Barnard College, Columbia University. Recently, he held the titles of Visiting Professor at REMESO / Institute for Research on Migration, Ethnicity, and Society, Linköping University, Sweden, and Visiting Researcher at the University of the Arts, Helsinki Research Institute, Finland.

His books include *The Cinematic Mode of Production: Attention Economy and the Society of the Spectacle* (Dartmouth UP, 2006), *Acquiring Eyes: Philippine Visuality, Nationalist Struggle, and the World-Media System* (Ateneo de Manila UP, 2006), *The Message is Murder: Substrates of Computational Capital* (Pluto Press, 2017), and *The World Computer: Derivative Conditions of Racial Capitalism* (Duke UP, 2021). He is currently working on two book projects: *Economic Media* and *Collateral Genocide*.

finance, media, and racialized violence concatenated to cheapen life for capital investment. The essay further develops the concepts of "the derivative condition," "currencies of violence," and "network derivatives" to examine how computational infrastructures transform social and semiotic relations into instruments of risk, valuation, and financial exposure. M–I–M′ crystallizes the political economy of this emerging computational mode of production: a regime in which information functions simultaneously as means of production, labor product, fixed capital, and mechanism of accumulation. Actually existing AI consequently represents not an escape from capitalism's dependence on labor and life, but an intensified capacity to extract value from them.

Introduction

Let's begin with the series I introduced in *The Message is Murder* (2017) and developed in *The World Computer (2021)* as the dialectically emergent general formula for capital: M-I-M′ where M is "money," M′ is "more money," and I is "information." Here "I" replaces the "C" in Marx's M-C-M′ (Marx, 1973: 250; 1976: 247), where C is "commodity" — the commodity, recall, is the means by which capital gets from money to more money. The commodity also represents the domain of production that hides the theft of surplus value during the labor process. M-I-M′ asserts, in brief, that the general formula for capital has evolved because the techniques for the accretion of surplus value have evolved, and further, that computation is not incidental, but essential to the expansion of capital. Capital has entered a new period in which the commodity-form now appears as an instance of information — an informed object in the financialized data networks of racial capitalism. The production of information (to generate more information, and thus more money) is paramount — a new paradigm. If, at least since Hayek (1945), capital was always a computer, its formalization by and as computation, renders it — or reveals it via sublation as — informatic. The implications are corrosive.

Relatedly, and indeed *dialectically speaking*, it is becoming increasingly apparent that genocidal violence can function as a highly cost-effective, and potentially profitable, means of disrupting existing configurations of the status quo, generating volatility, and thereby producing new information from information, as well as generating demand for *more* information. The global financial and political investments associated with mass killing may thus be understood as a pronounced symptom and consequence of the expanded domain of production and reproduction in which virtually anything can be treated as information. The genocide in Gaza, in this regard, brings the racial dimension of

racial capital into particularly sharp relief— not that there should be any doubt about the "racialism" of capital (Robinson, 1983), or that capital ever operates independently of racial differentiation. It is the methods and modalities through which such violence is mediated and operationalized that require particular analytical attention.

Racial abstraction, and *the granularity possible therein*, has become one type of information, though an extremely significant and mutable one that is deeply embedded in the operating system of capital. Its origins, however, clearly predate formal computation. Since before the Balfour Declaration, Zionism has pursued a settler-colonial project in Palestine, encompassing the Nakba and subsequent decades of conflict, displacement, military confrontation, and the institutionalization of systems that have been characterized by scholars as forms of apartheid and ethnic cleansing. European and U.S. political and economic actors have, over this period, become increasingly involved in the processes of racial formation associated with the Israeli state and in processes of ethnic cleansing and dispossession affecting Palestinian and Arab populations (Khalidi, 2020; Mishra, 2025). However, the turn of events since October 2023 crystallizes a new stage in which racism becomes increasingly computational. Statements and policies emerging from the Israeli political establishment indicate the growing integration of computational systems, including AI-assisted applications such as Lavender, The Gospel (*Habsora*), and Where's Daddy?, into military and security practices. Reporting has also documented connections between these technological infrastructures and major technology and defense companies, including Google, Amazon, Microsoft, OpenAI, Palantir, Elbit Systems, Rafael Advanced Systems, and Corsight AI. At the same time, computational media such as *TikTok*, as well as social media platforms more generally, have increasingly become part of the informational and communicative infrastructure through which contemporary conflicts are represented, contested, and operationalized.

At a September 2025 briefing for conservative influencers at the Israeli consulate in New York, the Israeli prime minister described social media as an important "weapon" in contemporary conflict and discussed the prospective acquisition of TikTok by Oracle and a group of U.S.-led investors. In discussing X and its owner, Elon Musk, he likewise emphasized the importance of social-media communication and influencer networks in shaping public opinion. These developments suggest that content production and platform-mediated

communication have become increasingly integrated into the political and informational dimensions of contemporary warfare, extending the battlefield into the domain of computational media.

Beyond the role of top political leaders, the algorithmic moderation and differential visibility of content on major media platforms, including Instagram, have been widely documented, particularly in relation to representations of Palestine and in other contexts where corporate interests intersect with state power, including in Turkey. Sudan provides another significant example of the relationship between computational media, political violence, and geopolitical interests. Allegations concerning external support for armed actors involved in the genocide in Sudan, including support attributed to actors linked to the United Arab Emirates, further illustrate the ways in which large-scale violence may remain comparatively underrepresented within dominant media environments. These disparities in visibility raise broader questions about the extent to which computational infrastructures participate in the mediation, normalization, and legitimation of political violence — not incidentally but now as conditions of possibility.

Following the change in ownership, the U.S. version of TikTok has also undergone changes to its platform governance and content-moderation practices, while the platform's technical infrastructure enables forms of geolocation and the inference or classification of users according to demographic characteristics, including race, gender, and religion. Such developments indicate how computational systems can simultaneously organize the circulation of information and participate in the production of social categories. Genocide, mass killing, and their legitimation therefore operate within a broader informatic environment that affects both semiotic and political domains as conditions of production—conditions increasingly shaped by the investment and ideological imperatives of racial capitalist accumulation and the value form.

Though I will treat these examples in greater detail elsewhere, I include them here. They urgently speak to our question of living information. The present essay focuses on the latter, but readers must remember that what is said here about the productivity of information and informatic labor is, from the point of view of this analysis, inseparable from the financailized milieu that sustains the solvency of capital and that at bottom this solvency means investments in genocide.

I am looking at two contrasting representations of Gaza—one depicting the city in its current condition of catastrophic destruction, the other, released by the Israeli Prime Minister's Office presenting a speculative, computer-generated future of gleaming towers, efficient infrastructure, and orderly prosperity. Both of these images — of Gaza as it currently is, and of Gaza as Israel would like to imagine it — can be understood as "rendered" by AI. Each functions as a data visualization shaped by computational and financial logics. In the first representation, Gaza appears as a landscape of rubble, twisted metal, collapsed buildings, and devastated neighborhoods: a material record of genocidal violence. In the second, Gaza is reimagined as a hyper-modern coastal metropolis, complete with high-rise architecture, integrated green spaces, and advanced logistical networks. The juxtaposition of these two visions—one documenting obliteration, the other projecting an unreal future—reveals how visibility itself becomes a question of mediation: where and how artificial intelligence operates, and how computation organizes what can be seen, imagined, or invested in?

While it is important to understand the computational means by which such representations come into being—and to recognize that they are images among images—it is equally important to grasp the financial parameters that shape the data from which they emerge. Whether through conventional AI image generation or through AI-driven targeting systems and weapons deployment, meanings and the lives attached to them are placed on the market, and calculation organizes what appears. The devastated urban landscape of Gaza becomes legible as information, just as the speculative future city becomes a financial fantasy rendered through computational protocols. In both cases, matter is processed as information, and information becomes a site of investment. International capital flows into weapons systems, development schemes, and geopolitical projects, treating the world as information and leveraging that information as risk instruments. As the obscene investments surrounding genocide demonstrate, new information becomes collateral, collateral becomes credit, and credibility justifies further investment.

Both the actually existing landscape of Gaza—reduced to rubble by AI-assisted weaponry—and the speculative future cityscape are transformed by the medium of AI. Their expressions and ontologies are reconfigured as they become subject to computational rendering and forms of programming. Each representation is a "message" addressed to particular audiences, constituted by

the manipulation of information to generate further, predictably profitable information. Yet both also contain, if one knows how to listen, the cries of protest and agony inherent in their referents: one in the material devastation of a people, the other in the erasure and replacement of that devastation with an imagined future built atop dispossession. Both stand as testaments to atrocity.

Images, movements, landscapes, and much more are materially encoded and transformed by AI, altered according to its protocols and economic exigencies. The destroyed terrain of Gaza becomes a dataset, just as the imagined future metropolis becomes a computational projection. AI, as a communicative and financialized medium, remediates social relations and radically transforms the relation between image and referent. It dissolves former ontological guarantees and replaces them with probabilistic, n-dimensional mappings whose operations are largely untraceable. The lower-dimensional relations of earlier media regimes—where one might live on ancestral land or rely on indexical representation—are reconfigured within a higher-dimensional vector space that exercises command over them. Ownership, authorship, authority, rights, responsibilities, labor, and remuneration are all reorganized under this computational regime. In this transformed semiotic milieu, each rendering becomes a financial wager, a bet on the future utility of an asset whose market value is yet to be determined.

AI is a prime example of what in my work I have been calling *economic media — it has both a semiotic and a monetary register.* Like other media, it too is simultaneously expressive and bound by economic protocols, but as the most volatilizing medium extant it has become the most attractive investment on the planet, leaving little room for doubt regarding capital's current view on its prospective returns. Let's remember before moving on that corporate AI is not a thing, but rather an ecology composed of material and social relations that provides new access points for ownership, utility, labor and liquidity. It is the next stage in the emergence of financialized media ecology that has become inseparable from ecology.

With AI, McLuhan's observation in the chapter "The Medium is the Message," namely that "the content of any medium is always another medium," (1994: 8), was perhaps never more true. AI devours the archive. And it is setting out to devour the world, its cultures, its peoples, its pasts. As it becomes the general interface, AI should also be seen as another step towards media

"convergence," that is, towards the devouring convergence of all prior media into computation. Without going into more detail here, we grasp AI as a reconfiguration of reference: of "the referent" and "indexicality." Rather than the pairing of signifier to signified in the (reduced) notion of "the sign", or the supposed one to one mapping of three dimensions into two, as was the conceit of optics and photography, we have the rise of *n-dimensionality*, a new type of spatial mapping whose point-by-point calculus is a vector synthesis that is effectively untraceable from the result. However, having devoured the archive, AI can simulate its objects in most if not yet all of its prior media, arriving at renderings that seem familiar though arrived at by unfamiliar means. Sometimes it is difficult to discern its operations. Robotics promises that this capacity for simulation will soon include embodiment. With the development of n-dimensional capacities, the planet and the cosmos become legible as information, encoded as information, volatilized by information, actionable as information, transformed *by and as* information.

This particular convergence potentially affects all prior nodes (objects) in media ecology, and posits the simulation and/or reformatting of any and all of them. The implication of this colonization of prior ontological formations by information is that anything whatever is potentially a site of production and arbitrage. As has been keenly observed (Crawford, Hao), AI places vast demands on planetary ecology. As part of media ecology and general ecology, AI is an abbreviation for an economic, semiotic, political, cognitive, ecological, computational, thermodynamic, material agenda that would reconfigure global energy production and distribution, nation states and governance, water and land usage, the psyche, intelligence, the the planetary mine, and, above all, capital's *exposure* to the extraction of surplus value as it battles the falling rate of profit by waging a counter revolution that endeavors to drive the value of living labor (and life) to zero. What is being sold as AGI, that is, artificial general intelligence, is in practice artificial genocidal intelligence. I will develop this point throughout my discussion.

Economic Media, Computation as Fixed Capital, Historicity of "Information"

Let's say that AI is the culminating moment in the history of economic media to date. What then is economic media? Economic media are media that manifest a semiotic component and a financial component. Their conceptualization posits and presupposes the increasing coherence among the domains of semiotics and valuation, production and circulation, information and computation. They organize meaning and economy (use-value and exchange-value), and their emergence as command-control utilities signals a subsumption (formal and increasingly real) of semiotics by racial capitalism. Where with the Gutenberg press, painting, printing and photography it may have once sounded out of place to insist upon attention to the economic logic underpinning a novel or a picture (though brilliant studies have been done of the novel (Moretti, 2013; Gallagher, 2006; Watt, 1957; Hughes and Lund, 1991) and patronage (Kempers, 1987; Haskell, 1963)), with cinema, television and computing, the econometrics begin to exercise ever greater powers of overdetermination with respect to the ostensible "content." As I wrote already long ago, "attention economy" is developed with cinema bringing the industrial revolution to the eye (Beller, 1994; 2006). With the famous Hollywood system and with cinema in general, economics begin to organize the imagination and thus more of the world by means of cinematic mode of production. Through feedback loops with capital economic considerations structured images, the forms of reception and the social effects of cinema's so-called content — effects that media companies set out to capture with increasing self-conscious purpose. Attention economy included spin-offs, sequels, fashion, gestures, life-styles, the increasing valuation of artworks, celebrity etc. But the social penetration and generalization of these predecessor media ecologies, that went through another layer of development with the rise of the internet powered screen, pales in comparison to social media and AI.

Computing swallows prior media. With the rise of the web, and especially with the rise of social media and its injunction to produce content, prior media, as well as new media creations, are (*re-)platformed on binary*, making it clear that *all messages are subject to both semiotic protocols and economic protocols*, and furthermore, that such *double registration* is the condition of nearly all meaning making. Like microplastics, air pollution, and climate change, global digital media forms bearing both sign value and economic value are always incorporated in the circuits of meaning-making to a greater or lesser degree, no matter how far one thinks one may be from an actual computer. Life activity responds to

computational states and creates state changes in computers. What's more, and this is a key point, *digital media are hosted on platforms that compose fixed capital*. Fixed capital seeks its returns.

That ambient interface with fixed capital composes the reality (such that it is) that underpins our watching, our writing, our image making, our semiotic interfaces with the rest of the world. One major problem for nearly all bodies politic is that we know how to work the semiotic layer of our media (we can say or image or imagine whatever we think we want) but we don't know how to work the economic layer — we have little awareness and even less control over the value creation and distribution consequent from our communications and dreams, including little control over the economic protocols pertaining to our cognitive and kinesthetic meaning making. Admitting this inability is to confess that ultimately, we have little control over the economic consequences of our thoughts, aspirations, movements, and work. For the most part, *Homo sapiens* doesn't know what *homo economicus* is doing.

Communicative media have a semiotic layer and an economic layer — in binding semiosis to finance today, they transmit and archive messages as they aggregate value for the lords of racial capitalism. This description of the dominant dispensation of planetary mediation is what I call *computational racial capitalism*. It functions with ostensible content indifference, the only criteria being that information, as a financialized form, must seek and ultimately provide a profitable return. If once stated this characterization seems incontrovertible, that is because it *is* the dominant business model: capital extracts information with small remuneration (or no remuneration) paid increasingly in platform currencies — likes, follows, maps, company scrips — in order to optimize its network in the network of networks. The implications of this re-organization of social intelligence and thus also global production by such a virtual machine — a world computer — have not been adequately reckoned.

Among the consequences is nothing short of a planetary reprotocolization of semiosis, and thus also of social production and reproduction, and thus the *experience* of planetary life. The subsumption of prior media by computation implies the subsumption of global semiosis by information. Here too we must pause briefly. Just as we see that computational infrastructure is fixed capital and therefore requires a return, we must understand that *information itself is not a transhistorical category but a product of a dialectics of extraction — a form of capital —*

enabling, in a word, the conversion of qualities into quantities, of concrete elements into number. It comes to the world in historical time at a certain point in the development of the value form and the scaling of racial capital; it emerges as a *heuristic* with which to conceive the world and a *pragmatics* with which to treat it. It extends quantification into the tissue of all things. Such translations of entropic forces (noise) into information render the word calculable in an increasingly granular and probabilistic way. This elaboration of a digital medium beyond mere money (that is also an extension of the metrics of money) for the active conversion of qualities into quantities ("Digital Culture 2"), follows upon *a first regime of digitization imposed by the commodity form in the form of exchange value,* ("Digital Culture 1;" Beller, 2021)*, in which qualities were indexed to number as price.* Numeracy increasingly integrates, vertically and horizontally, material forms with price and probability (financialization), while allowing for the assemblage of new types of assets: digital objects (Hui, 2016), etc. The global production of the world as informatic analytically and practically dissolves the continuities of prior Earth-based ontologies of world and cosmos and renders their *bits* subject to a global calculus — that of computational racial capitalism. Anyone running one of these calculators we call computers, and anyone feeding one of the calculators with content of any sort, is involved in disassembling and reassembling the world — creating order and entropy — and is doing so in an interface with capital — a deterritorialized factory. And capital must become profitable if it is to remain capital.

Even if (to preserve the virtue of science) one wanted to think that the rise of information and the datification of the cosmos were separate from the history of racial capitalism (I do not), such a one cannot fail to recognize that the recording, processing and archiving of information is a capital investment made in anticipation of returns, that is, of interest. Ultimately, the interest of information is money. From finance to pharmacology, from agriculture to astrophysics, all that data is fixed capital, and all interests therein are (or aspire to be) converted into bank interest. With this result in mind, and understanding that information is not one thing but an historically emergent modality of treating the universe that has grown increasingly complex in synch with the imperial complexity of capital, then the desire to see informatics as separable from the history of capital seems ideological, idealistic, naive, reactionary.

As soon as we understand information as *method* — a method of converting any interest whatever into interest for capital— and as requiring a material substrate, we may clearly apprehend the *platforming* of information as *expressions of fixed capital*. We also see that the massive investment in AI indicates capital's perception of a way to combat the falling rate of profit by moving more expeditiously from M to M′ (money to more money). Capital moves to sectors with the highest rates of return. Rates of return are inexorably linked to the ratio of invested capital to surplus value extraction by whatever means — direct exploitation or leveraged financial mediation. The higher the rate of return, the greater the degree of extraction — or, in the common case of speculation, of perceived future extraction. Datafication, the conversion of any quality whatever into quantities, is the means to current capital expansion. It indicates that the movement from M-M′, expressed here in what Marx called capital's "concise style," is mediated by information: thus, we write M-I-M′. This formula, as we have said, takes the place of Marx's general formula M-C-M′.

Living Information

Just as M-C-M′ signals that it is in the *production* of Commodities that the dissymmetrical exchange with labor takes place (classically, through the wage — the payment of which converts quality to number and, by leveraging the worker's precarity, extracts unpaid labor as surplus value —we are bound to recognize a similar if dialectically more advanced form of dissymmetrical exchange in the *manufacture* of information. Here ghost work, screen interfaces, the strip-mining of attention and libido, carceral time, becoming a target, all come to mind as new types of labor. "Workers," or whom I now think of as "*informatic laborers*," who are sources of "*living information*," *create state changes in discrete state machines* instead of simply informing matter as was the case with classical commodity production in which workers gave form to ordinary matter in exchange for a wage. With the Internet of Everything, *all state changes including* those of our old-school material creations such as cars, oil, and iPhones are multiply (n-dimensionally) indexed as information and may thus also serve as interfaces — sites for the accumulation of information. Information as distributed machine-form, as command-control logistical utility, as attentional input, and as fixed capital, networkologically *sublates* the production of mere commodities (though such objectifying partitions are still operable and, of course, fetishize-able).

With the emergence of the superset "information" from the set of all commodities comes the further abstraction of labor and of what Marx conceived as "living labor" as the scene of surplus value extraction. From labor that informed matter in more conventional ways, the expanded form of labor becomes, as I have previously stated, *informatic labor*, while the performance of this informatic labor is, in fact, *living information.* The embodied form of informatic labor, performed as living information in any interface whatever contributes dissymmetrically to capital's proprietary information by converting I to I′ — which is to say, by creating a (machine) state change — thus M-I-M′ becomes the general formula for capital because surplus value is aggregated as information — new information. What the informatic laborer receives in return for provisioning capital with new form is dependent upon their position in the network (their information profile) and is continuously re-negotiated.

Though abstract, none of these changes nor their consequences remain fully abstract. Put another way, the abstraction process is material, lived, feeds bank accounts, and also maims and kills. Indicated by this shift are tremendous transformations in the forms of labor, rule of law, commodity-objects, jobs, money and currencies, assets, risk management tools, social differences, network derivatives, options, futures and the very space and practice of politics. This latter is difficult to emphasize enough, since the older metaphysics of outmoded media regimes are laid to waste (truth, representation, objectivity, linear time, etc.). The dissolution of ontological guarantees, the compounding of the discriminatory social differences cultivated and institutionalized over time by racial capitalism, the institutionalization and normalization of genocide, along with carcerality, policing, and the racism, gender essentialism, and rampant psychosis that accompanies these strategies of media war, suddenly appear as among the most profitable ways of producing information. As Helga Tawil-Souri writes in "Bitificaton and the Gaza Genocide," "Palestinians and *Palestine* are continuously disassembled, reassembled, and processed into digital data, whether for the purposes of killing or media consumption" (2026: 13). This bitification is paradigmatic and indeed serves as the linchpin of *the computational mode of production.* Depending upon one's position in the network — where position is indexed to one's access to the social product, one's value — the volatilization of everyday life by bitification has varying consequences, even though it renders everything once solid as a risk instrument, a bet on the future. The further down

the network hierarchy one goes, the more precarious the future. In plain speech: the global economy has come to presuppose genocide as its fundamental underlier. Not only is genocide to be grasped as the other side, the Janus face of actually existing AI, it is what investors are investing in. Social categories become lived abstractions. And pathways to further pulverization.

We glean from the profiteering inherent in today's volatility-producing computational renders, including their programming of ecological disaster and planetary war, the now irreconcilable distance between information and anything like Justice or Truth. Even the version of this essay that you are reading has been redacted by (fear of) AI, particularly on the subject of Palestine, to comply with this journal's legal concerns, especially in Germany (the original version will be made available eventually). It is not only that these ideals are no longer practically available. In superseding prior modes of production, computation has shattered the metaphysics that set forth these old school imaginaries. With the volatilization of enlightenment rationality by information, the old metaphysics no longer pertain. The mode of production along with the material world from which they emerged has been vitiated. As already discerned in the *Dialectic of Enlightenment*, the end of reason is unreason — *that* is its rationality. For better or for worse, the struggle with the rationality of information and the racial capital for which it currently stands cannot be waged in classical political terms any longer. Alas, justice and truth belong to a vestigial metaphysics. They have been ushered out of history by informatics and the granularization of capital logic into the atomic, the cognitive, the bios, and the affective registers. There is no court of appeals, and no non-cybernetic species, and no God but power, at least in computational racial capital's philosophy and practice. Nor even, I think, can the *utopian* results proposed by the aspirational strains of those classical and enlightenment ideals be achieved through antiquated conceits of humanism or liberalism. In a world in which, as Cinxin Liu prophesies in *The Three Body Problem* (2014), *the very laws of physics are the results of an arena of cosmic war,* the struggle with informatics — the n-dimensional aspect of information — can *only* be a matter of politics. Recall that in that trilogy cosmic war is waged by collapsing dimensions, delimiting the understanding and field of action of entire civilizations. "Economic Media" "Computational Racial Capital" "The Computational Mode of Production," and the formula M-I-M′ provide conceptual-political entry points into this dimensionally exploded arena of unfolding disaster. In such a milieu, even God

is a risk instrument. And yet the power of belief may be the only thing that can save the species — but belief in what?

Information as Paradigmatic Site of Production, Arbitrage and Capital Expansion

It has been suggested that despite the apparent AI boom, model collapse is imminent. As AI starts to feed on content it itself has generated, the output degrades and rapidly becomes garbage. Our thesis here is that like capital — because it is capital — corporate AI always needs an outside. Without additional living information, it collapses. This living information is to be gathered at the borders of fixed capital — its interface with the living. It is at the network edges and interfaces that AI converts the world and grinds it into recomposable bits. The relentless encroachment on the lifeworld, the outside, the surround, is computational colonialism.

AI must continuously be fed information that it did not itself generate in order to avoid model collapse. We call this metabolic process, racial capital's interface with living information — the information it picks up from training sets, sessions, and other feedback loops with earthbound metabolisms too numerous to enumerate here. Living information feeds the dead information that constitutes the machine. Matteo Pasquinelli (2025) explains that the early theorist of *operaismo*, Romano Alquati, called such information — the information picked up from the worker's micro decisions and incorporated into the production regimens in the Olivetti computer factory "valorizing information." The processing and incorporation of valorizing information by capital converts both the existence and function of this information into capital, turning M into M′. One can observe the continuity here with Paolo Virno's notion of "virtuosity," in which the cognitive linguistic capacities of workers (and of people in general) are bent to the valorization of capital both in the factory and, as importantly, in the social factory (Virno, 2004). Upon consideration, I have selected the category "living information" as opposed to valorizing information because *valorization* of the capture of surplus labor/information requires the market and monetization to book the gain as profit — though admittedly (and importantly), fictitious capital can book unrealized gains. Though it is beyond the immediate scope of this essay, there is however, a case to be made for the category valorizing information, in as much as media forms in general, with the encroaching of economic media, have

become increasingly money-like and engagement with media forms adds value that can be booked — in, for example, share price.

For the moment, the market has expressed its full support for AI and its capacity to convert M to M'. Currently we have a planetary ecology of mining, energy, logistics, cognition, attention, affect, logistics, policing and warfare spear-headed by AI for the purpose of further entrenching AI into every aspect of social organization. The analysis presented here understands AI as the ongoing *formalization* of what I previously abduced as the world computer — the virtual machine constituted by what I have described variously as computational capital and computational racial capital, with the latter term serving as the more precise description (Beller, 2023). The current exponential growth of AI indicates that the world computer is becoming increasingly formalized and privatized. The market expectation is that such AI will produce the highest rates of return. How?

A desperate quest for content providers, that is, for living information, is unleashed, dwarfing even the former requirements of investments in the world wide web and social media. Content provision must now be understood as anything that feeds AI ecologies. The scaling requirements of the capital investment in AI requires ever more energy, rare earths, data centers, water, land, chip manufacturing (and all that goes along with that, including the ecological, environmental, and thermodynamic threats to local and indigenous communities around the world), and especially ever more and ever cheaper informatic labor, that is, living information to feed the already extracted dead information that constitutes both AI and fixed capital more generally. While the vast natural resources required to feed AI must themselves be treated as information in order to feed the prediction models that justify massive investment, the disappearance of jobs feared to be a consequence of AI expansion will not mean the disappearance of work. Interfaces with fixed capital will multiply and become ceaseless and the labor of *survival* will prevail as Peter Theil's idea of God reigns. While the new terms of work will be specified by a matrix of protocols and access contracts, the comparative security of the job contract (shitty as it was) will disappear, and the remuneration will, in general, be reduced to just enough liquidity for mere survival, unless, as is already true in many cases, death/killing can be made to turn a greater profit than can keeping certain folks alive.

As already indicated, it appears that the lowest cost and largest sources of living information (that is, the historical form in which capital can exercise the

most leverage against the living and take the largest proportion of life in exchange for the least remuneration), can be found in surveillance, policing, border patrol, incarceration, encampment, forced displacement, military forays and genocide. Tragically, yet somehow unsurprisingly, the most value can be extracted from lives already cheapened by the histories of racism and colonialism. Their hours can be accessed for less cost, their information can be thieved with less responsibility. Whole lives and historical lands can be taken in exchange for providing an information profile legible to pattern recognition seamlessly feeding the deployment of bombs and drones. The updating of these formerly statist controls over the oppressed are intensifications of pre-existing strategies for the securitization of private property on the one hand, and of the volatility inducing media practices promulgated by "extremism" on social media on yet another. In affective terms it has become a cliché that extreme images drive traffic — social media platforms have long been monetizing their draw. Beyond the business of making the deals to arbitrage and privatize state functions and absorb large portions of national budgets, this aspect of AI by which people treated as information create state changes in discrete state machines operating as proxy state-formations addresses two markets — those involved in relations of policing, and those who must respond to the spectacle that such actions generate. The mostly unremunerated labor of survival — as minority, migrant, indigenous person, impoverished person, suspect, alien, refugee, "illegal," inmate or target — in a world already rendered volatile and precarious by the apoplexy that racial capitalism previously imposed through the century's long churn of the abstraction processes resultant from its relentless historical calculus, turns out to supply the cheapest source of living information. Lavender and Where's Daddy, feeding off the biotic struggles of the colonized, and used to profitably produce targets in Gaza, has now been repurposed as ELITE for use in the United States. Europe will soon be ready for its own version, as the European Parliament recently endorsed its own anti-immigration policies that include the forced deportation and the offshore encampment of unwanted humans, as part of its recapitalization scheme.

The general cheapening of life accomplished by AI makes already surplussed populations — lives made cheap for capital by colonialism, racism, imperialism and globalization — cheaper still. With brilliance and startling clarity, UN special rapporteur Francesca Albanese in her paper "From Economy

of Occupation to Economy of Genocide" (2025), shows that genocide has become a business model — an economy. Increasingly it is *the* business model. And in every dimension, it is now being driven by AI.

While it may seem that only psychotic subjects could invent, manage and run such businesses, no rational person could deny that we have many such subjects in our midst. There may even be some reading this essay. These avatars of the world computer, evolving in the toxic data swamp of the racial capitalist matrix embody just the right combination of pathologies to constellate the vectoral tendencies of racial capital while remaining content indifferent when required. That's their heritage: Just the right mix of entitlement, delusions of grandeur, contempt for their fellows, sadism, objectification, and narcissism — in short, an infinite capacity to deny the existence of the others whom they make suffer and whose suffering is the precondition of their very sense of self, their *jouissance*. Sociogenetically these features provide a recipe for success (whiteness, maleness, capital), and are expressed as the righteous imposition of rigid hierarchy of social difference along the vectors of race, gender, religion, and nation, and increasingly along the invisual partitions of the new racisms possible in the patterns discerned by machine learning (Munster, 2025; Phan and Wark, 2021). This psychopathology in which the taking of others as objects in accord with one's desires is the means to constituting a self has long been the recipe for dominant agency in racial capitalism. These psychotic subjects depend upon the concerted and indeed hysterical maintenance of a profound ignorance of whatever is beyond the reified thresholds set out by monetary concerns, and, when, as often occurs, a principle that was invoked in one moment needs to be denied in the next, a moment-to-moment repression of rational contradictions.

Already well described through the characterization of O'Brien in Orwell's media-theoretical classic *1984*, this psyche, formed by newspeak and doublethink, and consistent with the agency behind AI, is the consequences of the history of slavery, conquest, settler colonialism, colonialism, imperialism and globalization reaching for immortality and totalitarianism through the calculated reduction of cognitive capacities by means of the promulgation of images and the expropriation of natural languages. In *1984* as in 2026, language is increasingly a mechanic extension, and the sustaining economic terms of society are to be understood as enabled in and by mediological paradigms. Totalitarianism of this kind is not possible without mass media, the screen, and the thought police, and

indeed, none of what the world is trying to live through right now is possible without capital. Today's totalitarian emergence entails the development of both semantic and monetary media and their ultimate convergence as economic media. Orwell (1961 (1949)) understood well that war was a means for the state to destroy the social surplus, and legitimate itself through an imposition of ignorance capable of checkmating thought. Recall the Party slogans: "War is Peace," "Freedom is Slavery," "Ignorance is Strength." He discerned the economic (il)logic of totalitarianism in the promulgation of stupidity and saw that it required the destruction of cognitive powers, solidarity, and indeed the metaphysics of the purportedly human, including logic and love. However, his analysis of extreme alienation along with the loss of the capacity to register contradiction, did not extend to the critique of media capital that material conditions today demand. The convergence of the ostensibly separate processes of representation and financial accounting is ultimately what vitiates the liberal individual psyche, dis-mediates traditional social bonds, and alters the conditions of social production and reproduction. Where once the individual as "the modern subject" was the privileged interface with capital's operating system, this general form of the interface with capital is no longer paradigmatic. Meanwhile capital's treatment of people as objects and as decomposable objects intensifies. As well, dividuals emerge as network artifacts, locally engaged in ways that do not translate from network to network. This is largely due to the fragmentation wrought by an overlay of qualitatively distinct networks.

The money forms and financial instruments that provide security and liquidity develop alongside writing, photography, cinema, videography and digital media should therefore be seen as bound to specific techno-mediological regimes that are endemic to subject and object making capacities inherent in the history of racial capital. Money is a *medium* and in as much as the form of labor has changed with technical and financial "advances," so too have forms of remuneration and currency. The wage for example, allows for the extraction of surplus value while assigning a market value to the worker. The racist image devalues a people while assigning a categorical ontology to a "them" or an "us" that in a circuit of violence cheapens their/our labor. The mediological constitution of agents and objects is driven by capital, overcoding prior ontological conditions. As racial abstraction, gender abstraction, and the abstraction of social differences becomes quantified and informatic, AI, which

depends upon continuous extraction and the assignation of probabilistic referentiality and is currently the largest financial endeavor in the world — becomes literally the *thought* of capital — the automation of its theory and *practice*.

Where belief, stoked by forces ostensibly outside the economic domain, once seemed to "naturally" follow the exigencies of capital, as with, for example, Humanism — later shown to disguise a constitutional racism as universalism— or, was carefully orchestrated in propaganda, pedagogy and early broadcast media as with Hitlerism, Americanism, treatises on the lazy native, and images of white supremacy, it is now quite explicitly and thus also most cynically a political-economic wager. With the full colonization of the sensorium by economic media, belief in a solid reality founded on fixed metaphysics, has melted into the instability of attentional capacities and then evaporated into the informatics driving the shape of consciousness and organization of metabolism. *Volatility* structures the metapragmatics of semiotics and converts meaning making into a probabilistic bet on a spread of outcomes — a wager.

In the war for labor, now attention, cognition and metabolism — what Neferti Tadiar calls "life-time" (2022) — all inputs are posited as living information to feed M-I-M.' These inputs are essential to the operation of the means of production as both fixed and variable capital. Just as the category of labor under capital requires the concept of living labor as distinct from dead labor to designate incoming contributions to the accumulation of dead labor, it now requires for informatic labor the category of living information as distinct from dead information to designate incoming contributions to the accumulation of dead information which is to say dead labor, which is to say capital. The accumulation drives the increase that feeds the vast automaton, which is to say AI, or what Pasquinelli understands as "the automation of automation."

Taking the slight revision of Bateson's definition of information that I made in *The Message is Murder* — information as a difference that makes a *social* difference — we now see that what I there called "substrates of computational capital," in other words we, the people of earth, are universally posited as sources of difference to feed the world computer's state changes. And, as we have observed, the dispossessed become at once the greatest source of such information because, like the workers before them, they/we are the greatest source of cheap information, stripped as they/we are of enfranchisement, and also the greatest threat to capital, to *racial* capital, that has become fully informatic. The

masses are at once the source of value for the valorization of computational capital and its biggest threat. Extending the injunctions of financialization to formalize practices of risk — itself a counter-revolutionary development of derivative techniques for the profitable management of the threats to capital, informationalization (or bitification) becomes essential to the profitable reconciliation of capital's books. Investment goes to the technologies that can most profitably exploit this contradiction between the organizational forces of capital and the forces of its demise, by extracting as much as possible for as little as possible cost, that is, maximum extraction for minimum remuneration. Whether as a worker in the factory or as an informer in the computational milieu, the laborer and now informatic laborer must also navigate this contradiction from the position of precarity — a binary of risk and reward operative at every point on the network. How to remain solvent enough to pay for a life in a world of vicious extractivism? How to survive when the decimation of everyday life requires victims to acquire more information to know how to survive.

Here too, in the cruel optimism of the digital, one sees the synthesis of AI and the military industrial complex — and we should add the role of the dollar as a global currency. For these technologies (and I include the medium of money and its compute) have together cheapened life around the world so that capital could perform the magic of arbitrage. "Innovation" that allows for the production of cheaper goods that can take over a market must be understood not only in terms of technological development, but in terms of getting more value out of workers for less cost. Capital can do more to people already stripped of rights and a social safety net. Unlike the technologists, we therefore include in our analysis of capital's innovations, the development of capacities to discount life, as did slavery, the plantation system, indentured servitude, the coolie system, the rules of apartheid, the reservation, and countless other strategies before it and current today, for the devaluation of being and thus for reducing the cost of labor for racial capital. Remember too that the cheapening of workers in one sector devalues all workers who must compete with those who have been devalued. Such a conflation of technical prowess and brute oppression not only corresponds to the history of capital but helps to strip the actual drivers of "innovation" from its self-mythologization as sheer technical advance. Racialization as much as computation. Borders and dirty wars as much as algorithms and logistics. The emerging "invisuality" of some forms of racism, in which pattern recognition in

effect creates its own racial categories not necessarily legible phenotypically, only further validates for capital the historical devaluations of racialized people(s). The compilation, aggregation, institutionalization and textualization of previously encoded racist practices informs pattern recognition such that actionable groups or targets may be categorized no longer only according to phenotype, blood, or DNA, but merely by n-dimensional edge detection that machine learning alone is capable of recognizing in its relentless pursuit and production of difference.

I think of this ongoing devaluation of the living as an "arbitrage on life." The imperative to reduce the costs of production of commodities requires the ability to reduce the cost of the production and reproduction of life, and thus of living labor/living information, and thus of its remuneration by capital. As with wage labor, consumer society, and social media, capital self-organizes against people's quests for freedom, in order to farm planetary desires for liberation and revolutionary transformation. Rather than elaborating here on the many expanding dimensions of the capacity of capital for theft that have, over time, compounded historical violence to produce interest for capital, I will conclude here by reiterating some of the new terms that at once signal the displacement or sublation of prior concepts by the onset of what I have called computational racial capitalism and the rise of M-I-M'.

The Derivative Condition, Network Derivatives

Adding to computational racial capitalism, the world computer, the arbitrage on life, informatic labor and living information we have the following: Derivative conditions (or "the derivative condition"), in which socio-semiotic outputs are constituted as wagers on the value form — risk instruments with which to manage and if possible benefit from the volatility of capital. From this perspective, an Instagram post or a database and everything in between can be seen as a risk instrument. The term economic media invokes and conceptually formalizes the *convergence of monetary and semiotic media,* always perhaps present (despite the seeming differences between say literature and stocks), but money and meaning are increasingly imbricated to the point where both registers are nearly always operative. Of course, the culmination of this convergence is in computation which manages both communication and finance — and now in AI which is nothing less than the thinking of capital. Within the framework of economic media we see the emergence of what I think of as Currencies of

Violence, that is currencies that are founded on formally or less formally institutionalized violence. One can see how the US dollar is a clear example of this, where the income derived from *seigniorage*, that is, from the act of issuance, depends upon the dollarization of planetary economics and its maintenance by means of the military industrial complex (Linares, 2026). I have used the term to also describe the horrifying significance of feminicide by drug cartels in Mexico, who use their capacity for violence to impose a regime of truth that confers recognition to the socio-semiotic practices of the cartels and enable the *conversions of messages violently written on the bodies of women* through torture, rape, scarification and incineration, *into harder currencies* exchangeable for goods in the "real economy" (Beller, 2023). Of course, and indeed paradigmatically, the value of the multiple issuances under brand Israel is deeply related for the networked capacity to execute violence in conducting and legitimating a genocide, but that is the topic of a different book.

Given these formations of currency and credit from social-material relations, we also observe the constitution of *new asset classes*, network derivatives, that give exposure to the extractive powers of economic media. Network derivatives rely on the implantation of technical networks, including and especially computing, in the bios. *Network architectures* constitute the *metapragmatics* of *network semiotics*. Machinic *interfaces* serve as *worksites: this includes, screens, keyboards, sensors, tokens, words, and more.* Basic network derivatives, such as cryptocurrencies, are quantifiable bets on the value of a network (as are, in this light, fiat currencies), but these derivatives can be synthetically composed as risk portfolios, in which the value of one token is tethered to a portfolio of other assets, and more complex financial instruments such as stable coins and even traditional stocks, etc., all of which can be incorporated into yet another layer of derivatives written on top of them. The general point is that *networked computational architectures* from financial systems to social media to cryptocurrencies to surveillance systems to AI *serve as forms of fixed capital* that provide exposure to networked value extraction (as informatic labor) in order to generate their returns.

Table of Concepts Endemic to the Sublation of M-C-M′ by M-I-M′

Economic Media — Semiotic Media plus Monetary Media (aka Money) — eg, social media, AI, USD. Semiotic signal and value signal carried on the same chain.

Computational Racial Capitalism — Capitalism, Racial Capitalism, Digitality.

The World Computer — The virtual machine composed by the globally distributed calculus of price signals and their derivatives.

Information — as the general form of the commodity, as fixed capital, as composable asset.

Informatic Labor — Labor, also "metabolic labor," also "life-times" (Tadiar).

Living Information — Living Labor provided at interfaces with fixed (informatic) capital, including screens, sensors, geolocation, etc.

Read/Write Ontologies — Dissolution and re-encoding of the essence of things by computation/AI. Read/Write Ontologies are broadly applied and potentially endlessly iterative. The term indicates the colonization (over-coding, re-coding) of prior ontologies and the reformatting of antecedent metaphysics. Eg., the violence and mutability of racism at the speed of light.

Derivative Conditions — Imperative for all denizens of computational racial capital to risk on information packets as value-positions on a spread of possible futures in a universe volatilized by the same.

Network Derivative — Exposure to the future value of a network

Network Superpositionality — Overlapping probabilistic wagers on a spread of possible futures resulting from the competing forms of future casting that computational racial capital (and its resistances) engage in.

Currencies of Violence —Economic media indexed to specific (often "innovative") strategies of extraction. Includes visual media, memes, utterances, conceits, ideologies, tokens, and monetary media.

Liquidity — Degree of Convertibility of a (quantified) form of economic media; command over life. In racial capitalism, liquidity is secured through the alienation and desperation resultant from the dissolution of ontological guarantees and its consequent dispossessions, that is by the radically dispossessed who must do/suffer (anything) to purchase their own survival.

Currencies of Historical Redemption — The non-alienation of production where wage-labor and informatic labor, etc., is replaced by currencies that do not sever the relationship to "ownership." The worker's product is abstracted but not extracted and is reconfigured as communal sense making via the tracking and assemblage of strands of stake in larger

socially connected objects, formerly understood as commodities and digital objects but becoming communal belonging(s). A Vertovian tapestry of stake emerges where the individual is bound up in the collective but not lost within it. These currencies amount to social derivatives written by the masses and providing the signers of this form a shared return. Multiple contributions to any given "object" can be indexed such that contributors retain rights/representation/legibility. The disenfranchised of the Earth will access the networks that provide them with the greatest degree of recognition for their contributions which also means that people work to build a collectively desirable shared future in a context of mutuality. Workers, migrants, the precarious will create these networks for themselves.

Communist Computing — The computational architecture capable of rendering currencies of historical redemption interoperable while developing their cooperative, non-capitalist capabilities. The computational cybernetic substrate necessary for species' survival.

References

Albanese, Francesca (2025) *From economy of occupation to economy of genocide (Report of the Special Rapporteur on the situation of human rights in the Palestinian territories occupied since 1967).* Human Rights Council, 59th session, A/HRC/59/23. Geneva: United Nations.

Beller, Jonathan (1994) "Cinema, capital of the twentieth century." *Postmodern Culture* 4(3): 1–32.

Beller, Jonathan (2006) *The cinematic mode of production: Attention economy and the society of the spectacle.* Hanover: University Press of New England.

Beller, Jonathan (2017) *The message is murder: Substrates of computational capital.* London: Pluto Press.

Beller, Jonathan (2021) *The world computer: Derivative conditions of racial capitalism.* Durham: Duke University Press.

Beller, Jonathan (2023) "Arbitrage on life, difference of the flesh: Racialization and colonial gender formation as algorithmic innovation." In: *The body in the field of tension between biopolitics and necropolitics.* [Publisher information forthcoming].

Crawford, Kate (2021) *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* New Haven: Yale University Press.

Gallagher, Catherine (2006) *The body economic: Life, death, and sensation in political economy and the Victorian novel.* Princeton: Princeton University Press.

Hao, Karen (2025) *Empire of AI: Dreams and nightmares in Sam Altman's OpenAI.* New York: Penguin Press.

Haskell, Francis (1963) *Patrons and painters: Art and society in Baroque Italy.* New Haven: Yale University Press.

Hayek, Friedrich August (1945) "The use of knowledge in society." *The American Economic Review* 35(4): 519–530.

Hughes, Linda K. and Michael Lund (1991) *The Victorian serial.* Charlottesville: University Press of Virginia.

Hui, Yuk (2016) *On the existence of digital objects.* Minneapolis: University of Minnesota Press.

Kempers, Bram (1987) *Painting, power and patronage: The rise of the professional artist in Renaissance Italy.* Translated from Dutch by B. Jackson. London: Penguin Books.

Khalidi, Rashid (2020) *The hundred years' war on Palestine: A history of settler colonialism and resistance, 1917–2017.* New York: Metropolitan Books / Henry Holt Company and Co.

Linares, Julio (2026) *Decolonizing Money: The Promise of abolishing the US dollar.* London: Pluto Press.

Liu, Cixin (2014) *The three body problem.* Translated from Chinese by Ken Liu. New York: Tor Books.

Marx, Karl (1973) *Grundrisse: Foundations of the critique of political economy (rough draft).* Translated from German by Martin Nicolaus. Harmondsworth: Penguin Books in association with New Left Review.

Marx, Karl (1976) [1990] *Capital: A critique of political economy, Volume 1.* Translated from German by Ben Fowkes. London: Penguin Books in association with New Left Review.

McKeon, Michael (1987) *The origins of the English novel, 1600–1740.* Baltimore: Johns Hopkins University Press.

McLuhan, Marshall (1994) *Understanding media: The extensions of man.* Critical edition. Cambridge, MA: MIT Press.

Mishra, Pankaj (2025) *The world after Gaza: A history of the present.* New York: Penguin Press.

Moretti, Franco (2013) *The bourgeois: Between history and literature.* London: Verso Books.

Munster, Anna (2025) *DeepAesthetics: Computational experience in a time of machine learning.* Durham: Duke University Press.

Orwell, George (1961) [1949] *Nineteen eighty-four.* New York: Signet Classics.

Pasquinelli, Matteo (2023) *The eye of the master: A social history of artificial intelligence.* London: Verso Books.

Phan, Thao and Scott Wark (2021) "Racial formations as data formations." *Big Data & Society* 8(2): 1–12. doi: 10.1177/20539517211046377.

Robinson, Cedric J. (1983) *Black Marxism: The making of the Black radical tradition.* London: Zed Books.

Steyerl, Hito (2025) *Medium hot: Images in the age of heat.* London: Verso Books.

Tadiar, Neferti X. M. (2022) *Remaindered life.* Durham: Duke University Press.

Tawil Souri, Helga (2026) "Bitification and the Gaza genocide." *Communication, Culture and Critique* 19(1): 13–24. doi: 10.1093/ccc/tcaf057.

Virno, Paolo (2004) *A grammar of the multitude: For an analysis of contemporary forms of life.* Translated from Italian by Isabella Bertoletti, James Cascaito and Andrea Casson. Los Angeles: Semiotext(e).

Watt, Ian (1957) *The rise of the novel: Studies in Defoe, Richardson and Fielding.* Berkeley: University of California Press.

AI as Oxymoron:
A Frankfurt School Critique of Intelligence, Information, and Aesthetic Rationality

Mary Caputi[1]

Abstract: This article examines artificial intelligence (AI) through the critical lens of Frankfurt School theory, questioning the apparent contradiction embedded in the term "artificial intelligence" and its implications for human cognition, democracy, and emancipation. Drawing primarily on the work of Theodor Adorno, Max Horkheimer, and Herbert Marcuse, the article argues that AI's growing capacity to gather, synthesize, and reproduce information may simultaneously contribute to the erosion of critical thinking, imagination, and the capacity to challenge established forms of knowledge. AI is situated within the contemporary extension of the "culture industry" and the "administered society," in which technological efficiency, instrumental rationality, and convenience can encourage conformity and intellectual passivity. The article distinguishes algorithmic processing from the mediated, dialectical character of human thought, emphasizing that genuine intelligence emerges through the continually changing relationship between subject, object, and lived social reality rather than through the automatic production of predetermined responses. At the same time, the analysis rejects a wholly pessimistic interpretation of technology. Drawing on Marcuse's concept of "aesthetic reasoning," it identifies imagination, negativity, and the capacity to envision alternatives to existing

[1] **Mary Caputi** is Professor Emerita of Political Theory at California State University, Long Beach, where she taught political theory, feminist thought, and critical thinking for thirty years. Her scholarship spans feminist political theory, critical theory, American culture, and the politics of power and resistance. Her most recent publication is the Research Handbook on Feminist Political Thought (Edward Elgar, 2024), co-edited with Patricia Moynagh. She is the author of Slow Culture and the American Dream: A Slow and Curvy Philosophy for the 21st Century (Lexington Books, 2022), Feminism and Power: The Need for Critical Theory (Lexington Books, 2013), A Kinder, Gentler America: Melancholia and the Mythical 1950s (University of Minnesota Press, 2005), and Voluptuous Yearnings: A Feminist Theory of the Obscene (Rowman & Littlefield, 1994). With Amirhosein Khandizaji, she co-authored David Riesman and Critical Theory: Autonomy versus Emancipation (Palgrave Macmillan, 2021).

conditions as enduring sources of resistance to technological and social domination. Aesthetic reasoning, understood as fluid, oppositional, and rooted in concrete experience, offers a means of preserving human autonomy and critical distance in an increasingly technologized world. The article concludes by considering contemporary forms of resistance to digital saturation, particularly among young people who seek greater connection with nature, art, community, and offline experience. Such developments suggest that the administered society has not achieved total domination and that the possibility of a "new sensibility" remains alive. AI, therefore, presents not only a technological challenge but also a critical question concerning what it means to think, imagine, and remain human.

Introduction

The developing world of Artificial Intelligence (AI) occupies a large presence in today's public sphere, dominating many conversations and orienting many decisions. The cutting-edge impact of its extensive reach represents the proverbial double-edged sword whose beneficial attributes and promises cannot be severed from its more worrisome qualities in a supposedly democratic, free-thinking society (Christian, 2020; Coeckelbergh, 2022; Zuboff, 2020). A multi-tasking proxy for the human brain, it functions as a virtual research assistant, a fact-gathering aid, a problem-solver, advisor, and mentor that offers advice thanks to its ability to mimic the thought process and imitate human behavior. It uses algorithms and synthesizes data in order to operate as a font of useful knowledge, demonstrating not only impressive tracking skills but also interpretive powers and the sophistication of highly educated, expert minds. For many, it represents the future and is pregnant with life-enhancing possibilities.

Yet AI also carries pejorative resonances, bringing with it attributes that are anything but life-enhancing. One obvious downside is that it eliminates jobs and alters traditional career paths, displacing persons who have long been employed while thwarting the hopes and dreams of those just entering the workforce. Indeed, for many either looking forward to retirement or seeking job security, AI arouses fear and a sense of dread as it represents the worst in precarity capitalism, replacing people with programmed computers, robots, Alexa, bots, and other technological creations capable of performing tasks without the standard job benefits needed by human beings. For many, AI thus spells insecurity. "Americans are anxious about their jobs and specifically about whether AI is coming for them," explains Geoff Bennett of the *PBS NewsHour* (*The PBS NewsHour*, April 10, 2026).

Yet job insecurity and the travails of economic precarity are not the only point of contention. For all of AI's many benefits, critics note the colonizing effect this technology has on our minds, fearing that its deleterious influence in fact outweighs its virtues. The worry is that AI encourages intellectual torpor and a regrettable erosion of critical thinking skills, producing a world wherein falsely contrived information is virtually indistinguishable from what is real, and wherein consumers not only fail to discern the difference but lose the drive to investigate. This inability to distinguish what is carefully argued from what is superficial, to call out "fake news" and other artificial confections, carries worrisome possibilities for democratic societies founded on dissent and free, original thinking. For while it might appear to encourage a healthy populism that allows marginalized voices to be heard, AI also facilitates "narrowcasting," "echo chambers," and "information bubbles," selective social networks shared by like-minded persons unreceptive to alternative worldviews (McNish and Galliott, 2020). Despite its potential to foster pluralism and debate, it can narrow the playing field and make users all too comfortable in their mindset. In this way, AI poses a danger to democracy, thwarting the critical thinking skills of citizens who believe they are socially engaged when in fact they are not.

The writings of Frankfurt School scholars help us probe this issue, since several of them expressed similar concerns going back to the mid-twentieth century. Their insights into the "administered society," the loss of critical thinking skills and subsequent acquiescence in a "one-dimensional" culture help us analyze the dangers of AI, whose very title seems an oxymoron: how can intelligence be artificial? Thus, while some things have changed drastically since critical theorists wrote their influential texts, their writings still speak directly to us despite the intervening decades.

Culled, Not Queried

Given the ubiquity of AI, it is useful to ponder its presence in terms of the "culture industry" about which Theodor Adorno, Herbert Marcuse and others wrote as far back as the 1940s (Adorno, 1991; Horkheimer and Adorno, [1947] 2002; Marcuse, 1964). Placing the word "culture" alongside "industry" gives us pause in the same way that the term "artificial intelligence" invites scrutiny. Seemingly, the two words inhabit radically different spheres and thus don't belong together. Yet as early as 1947, Adorno and Max Horkheimer deplored the

alignment of commercial and cultural interests which caused an atrophying of the public's critical insights, an atrophying of which people were often not even aware. Film, music, television, and other art forms exerted a dangerously conforming influence on the public, for "culture" no longer performed its necessary function of encouraging the deep, oppositional reflection essential to democratic societies. Culture was now dominated by "the total power of capital" which itself answered to the needs on commercialism and industry. "Technical rationality today is the rationality of domination…the compulsive character of society alienated from itself" these authors write in *Dialectic of Enlightenment* (Horkheimer and Adorno, [1947] 2002: 94-94). Instead of giving rise to free expression, the culture industry produces an "administered society" that only appears to be founded on democratic principles. In fact, it is "engulfed by an insatiable uniformity" that forever reproduces itself (Horkheimer and Adorno, [1947] 2002: 97).

Surely the anaesthetizing impact of the administered society is potentially reproduced by AI, whose primary task is to do our thinking for us. It must be said, of course, that the alacrity of AI's fact-gathering skills is often beneficial, and it cannot be denied that it aids immensely in such fields as medicine, environmental science, and political and military intelligence. Advances are being made thanks to AI that improve the human condition and thus, as per Hannah Arendt, make us feel more at home in the world (Arendt, [1958] 1959). Yet the cautioning words of Frankfurt School scholars still need to be heeded, since the inroads made by *artificial* intelligence have surely done damage to the *native* intelligence of those who rely heavily on its quick-and-easy approach to thinking. An educated, intellectually curious, thoughtful public is essential to a democratic society, and the ability to transform native intelligence into something computer-generated should surely give us pause.

The potential erasure of this ability to think critically, relying instead on hard-working screens and talkative, information-gathering devices, may well be the latest frontier in the "administered society" that promises a more practical, more efficient, more streamlined life while delivering not progress, but regression. As a critic of the administered society, Adorno looks askance at such promises which he labels the tools of a profit-oriented, dehumanizing system. It is a system that teaches us to treat one another instrumentally, as units of exchange subject to neoliberal market ideology, not as human beings who are

ends in themselves. "The practical orders of life, while purporting to benefit [wo/]man, serve in a profit economy to stunt human qualities, and the further they spread the more they sever everything tender," he writes (Adorno, 2005: 40-41). If AI benefits humanity while causing our critical thinking skills to atrophy, we at times struggle to distinguish its benefits from its harm, its progressive from its regressive impact. If our more humane, tender qualities are stunted even as benefits abound, can we hope for a more livable world, a place that does not tether "progress" to a loss of humanity? It seems that texts composed in the 1940s are still pertinent today.

"*Artificial* intelligence." As we look askance at the coupling of these two words, the writings of Frankfurt School scholars help us query whether intelligence can ever be truly external to human cognition, human experience, and the immanent situatedness of people. Those familiar with these authors' writings recognize immediately the jarring disconnect between "intelligence" derived from algorithm-fed computers and that which emanates from a critical, dialectical understanding of how human beings apprehend the world. For the world we live in – the world of everyday experience full of contradiction, inconsistencies, and puzzling realities – does not always conform to abstract ideas. The place where the rubber hits the road often makes trouble for erudite reflection while the latter is still needed to give meaning to the former. Those schooled in critical theory see immediately the opposition between the ready responses of AI culled from existing databases and Adorno's insistence that thinking is in fact *mediated*, fluid, a changing relationship between the knower and the known wherein the opposing poles of mind and matter inform and influence one another. Intelligence is "doubly mediated" by thought and materialism, ideas and society, not manufactured outside of us. In "Subject and Object," Adorno writes:

> The separation of subject and object is both real and illusory. True, because in the cognitive realm it serves to express the real separation, the dichotomy of the human condition, a coercive development. False, because the resulting separation must not be hypostatized, not magically transformed into an invariant…[T]he *pseudos* of the separation is manifested in their being mutually mediated – the object by the subject, and even more so, in different ways, the subject by the object (Adorno, 2000: 139).

Given this double mediation wherein subject and object are always distinct yet always conjoined, the thought process should be slow, deliberate, devoted to unpacking the received wisdom and distrusting the high regard for instrumentality, efficacy, and speed. It should never be automatic. While the prevalent culture industry that now proffers AI champions technological efficiency above all, we should scrupulously query efficiency and resist speed: the delivered responses are perhaps missing something, one-sided, lacking context, detail, and nuance. Algorithm-led reasoning necessarily assumes an instrumental approach to logic, thinking, and problem-solving, yet in so doing offers the end result as a finality which denies critical theory's insistence on the fluid, social reality of thought. While it can merge, synthesize, and rearrange existing patterns of knowledge in ways that propose alternatives, these alternatives base themselves on the data *already available* to AI's data banks and thus, at least in some cases, represent a reconfiguring of what is already given. The latter may well be thought-provoking and "intelligent" in its own right, yet the work of AI cannot approach the truly original, dialectical thinking of which socially situated human beings are capable. Adorno demurs: "If the focus of serious concern is with the efficacy to which it always appeals, the potential of its effect becomes twice as weighty… [consumers'] consciousness is further developed retrogressively" (Adorno, 1991, 105).

There is danger, then, in the widespread acceptance and even aggressive promotion of AI in the public sphere. The workplace, universities, industry, government and the military all make ample use of artificial intelligence today, and this is often a good thing. Yet the troubling aspects of this technology need sustained scrutiny. While voices of dissent are heard, some argue that insufficient critique has been undertaken in a world too eager to embrace the many conveniences that AI offers. As Bennet's interviewee, Josh Tryangiel, explains, AI "is not flawless" and does make mistakes. Yet "the AI industry has done a pretty lousy job of educating people about what it's good for…this is an important technological moment" (*The PBS Newshour*, April 10, 2026).

Frankfurt School scholars identify social acceptance, convention, and acquiescence as key players in both the popularity and the danger of AI's widespread existence, for by now it has gained such currency that, in some circles, its value is not even debated. The speed with which computer-generated

"thinking" can provide answers seems an unequivocal boon to many fields and professions. Yet the relationship between the human thought process and what is labeled "intelligence" risks being eclipsed by the impressively accommodating powers of technology: we lose our ability to gain distance on what is before us and to imagine something different, to query what is being offered and think beyond its limits. It is too easy to confuse the product of existing society, which Marcuse terms the "internal coherence" of convention, with the limits of what can or might be, since alternatives to the status quo are potentially vanquished by what offers itself as definitive truth.

This is especially true among young people who, having grown up AI-dependent, potentially cannot separate thinking, writing, and research-gathering from the multi-tasking of sophisticated technology. At the extreme, ideas for them come from machines, and thought is always pre-fabricated. The thought process has no life of its own apart from the apparatus that delivers it – the keyboard, the screen, the voice, the sophisticated platform that deploys AI models. All we need to do is cull, not think on our own. While some schools and universities demand that critical thinking be taught as a stand-alone subject, this requirement is not universally included in the standard curriculum. Having never encountered critical thinking as a subject of academic investigation, students have no reason to practice its skills nor apply it to their everyday lives.

Indeed, for life-long AI users, the thought process potentially reduces to that which can be delivered unmediated by the mind and unrefined by human introspection. Externally fabricated, its utilitarian outlook may foster brutality: "Matter-of-factness between people, doing away with all ideological ornamentation between them, has already become an ideology for treating people as things," Adorno opines (Adorno, 2005: 42). Marcuse similarly sees a loss of humanity in this inability to think beyond what is offered, to conceive something different, which spells a complete atrophying of the imagination and the possibilities that it brings forth. "The reality being reduced (or reducible) to its physical mathematical structure entailed that "truth" became defined by what could be measured or calculated," Marcuse writes "or by propositions which fulfilled these conditions" (Marcuse, 1989: 120). Viewed from this angle, AI delivers an updated, twenty-first century version of Marcuse's 1964 "one dimensional" existence artfully dressed up as "progress," having deftly coupled a loss of critical thinking skills with the promise of a more fulfilled, happier life.

If Tyrangiel is correct regarding AI's "lousy" job of educating us in this technologically important moment, this helps explain the ready acceptance and widespread acquiescence that our society evinces toward computerized thinking. The distinction between AI-generated answers and mediated, dialectical thought must be underscored more dramatically, allowing the perils of technology to emerge. Again, earlier texts by Frankfurt School scholars speak to the present, for Marcuse's one-dimensional society is precisely a society devoid of self-reflection and lacking critical distance on its own expression. In the 1960s, Marcuse was dismayed at how the people – "the organism" – were so conditioned to accept the mainstream culture that, often, they *instinctively* preferred the status quo over any form of rogue sensibility. While the Beat Generation and later hippies did present alternative, dissenting voices, the status quo which they opposed was by then deeply entrenched in the mainstream. "The organism is thus being preconditioned for the spontaneous acceptance of what is offered," Marcuse wrote in *One-Dimensional Man*. "Inasmuch as the greater liberty involves a contraction rather than extension and development of instinctual needs, it works *for* rather than *against* the status quo of general repression – one might speak of 'institutional desublimation" (Marcuse, 1964: 74).

If the reach of the administered state was already dangerously extensive by mid-twentieth-century America, attempting to colonize all forms of sublimated intellectual energy, how much more today when so many of us associate thinking with programmed devices. Indeed, the current commodification of knowledge and resultant intellectual torpor render the insights of the Frankfurt School even more pertinent to our social and cultural landscape. Against the backdrop of our damaged life, these authors' insights still seek out the trace of hope that resides in a colonizing, administered world. In this important technological moment, it behooves us to again find such traces, and to allow the insights of critical theory to supplant the "lousy job" of the AI industry.

"Another Reason, Another Sensibility"

Marcuse's writings champion our instinctual survival of capitalist cooptation, the fact that, despite the overwhelming influence of the damaging administered society, human beings retain the imagination to desire something different. While the pessimism of his *One-Dimensional Man* decries the far-reaching scope of capitalism's grasp, other works by him point to the survival of

an instinctual life still unconditioned by the forces of oppression, a life that is not damaged, but that refuses the broadly accepted preconditioned culture based on repression (Marcuse, [1955] 1966). This is where the category of "aesthetic reasoning" comes into play, that form of cognition not tethered to repression but still able to imagine and indeed create a world beyond what is given. At the level of our basic drives, Marcuse maintains, we desire a world not dominated by the dehumanizing injustices of precarity capitalism, but one infused with humanistic values wherein exhausting labor is not our primary activity,

Indeed, despite the reign of the status quo, the instinctual human energies wanting to be free from manipulation remain. Marcuse argues that our ingrained affirmative eagerness to create a more humane society has not been utterly destroyed by the domination of precarity capitalism; colonized to a good degree, yes, but not utterly eclipsed. The human body forced into its role as "an instrument of labor," an entity mediated by market relations, still retains the instinctual memory of its purpose extending beyond the demands of labor (Marcuse, [1955] 1966), a purpose that includes leisure, beauty, and the importance of human connection. In *Eros and Civilization*, Marcuse insists that it remains possible to eliminate "not labor, but the organization of human existence into an instrument of labor" (Marcuse [1955], 1966, 155) such that each individual lives not merely to earn, "but as a subject of self-realization" (Marcuse [1955], 1966: 210). This is the impulse of Eros as understood by Marcuse, an impulse which countervails the instrumentality treating people as things and prizing administered "truth" over deep reflection and dialectical reasoning. Eros, for Marcuse, represents an original human instinct shielded from the social reality of repression; it exists at the bedrock of our psyche unchanged by precarity capitalism. Although mediated thought insists that thinking and social existence constitute interconnected, interpenetrated poles, still this basic, pleasure-dominated instinct of the "organism" remains. "The memory of gratification is at the origin of all thinking," he writes, "and the impulse to recapture past gratification is the hidden driving power behind the process of thought" (Marcuse, [1955], 1966: 31).

At first blush, it might seem that "aesthetic reasoning" – a term also used by Adorno – represents yet another oxymoron, a term that once again reflects an internal contradiction not unlike "artificial intelligence" or "culture industry." For how can the realm of aesthetics, the realm of art and beauty, of sensual reality

and the primacy of the senses, be reconciled to the abstracted world of reason? The Western philosophical tradition offers numerous examples of thinkers who distrust sensual, affective reality and instead posit measured reasoning unsullied by emotion as a more rigorous and pristine form of cognition. In *Apology*, for instance, Socrates excoriates the poets for appealing to human emotions in ways that indulge the mind and detract from the perfection of virtue that only philosophy can advance. Poets, he argues, are like "seers and prophets who also say many fine things without any understanding of what they say…they thought themselves very wise men…which they were not" (Plato, 2002: 27). Yet Marcuse, like other critical theorists, insists that our aesthetic intuition, Eros, does not oppose reasoning but remains integral to it: it is, in fact, that which can be joined to philosophy and allow for a new, emancipatory society to emerge. Alloyed to the imagination and our ability to think beyond, aesthetic reasoning contains an inherently "negative" dimension that forces lived, everyday reality to query, inform, and reconstitute what the abstracted mind delivers.

As a mediated relationship between the mind and our situated, historical reality, human cognition thus registers how a given reality may falter, oppose itself, and contain internal contradictions; once we perceive this and resist settled ideas, we employ the aesthetic reasoning that is never instrumental but always fluid. Lived reality in all its opposition and subtlety allows the imagination to resist the fully colonizing grasp of the administered society, giving expression to the transformative, indeed revolutionary hope that aesthetic reasoning contains. For Marcuse, the aesthetic dimension "*transcends* its social determination and emancipates itself from the given universe of discourse," thereby drawing upon the instinctual life that, however buried, resists colonization by the administered society (Marcuse, 1977: 6). While the latter exerts a great amount of influence today in the form of information technology – "big data" – critical theory insists that a basic human instinct for dialectical reasoning has not been vanquished, and that the countercultural potential of mediated, oppositional reasoning still informs our understanding. "Only thinking could offer an escape," Adorno concurs, "and then only that thinking, the results of which are not prescribed" (Adorno, 1991: 200).

Technologically-generated thought thus does not quell the imagination's ability to force reason to work against itself, to see ingrained contradictions in the status quo that point to alternative meanings. The power of the imagination

remains, allowing at least the potential of a "new sensibility" to emerge (Marcuse, 1969). "Despite the generalized lack of revolutionary consciousness which Marcuse posits," writes Javier Sethness Castro, "he observes dialectically that, at their very heights, civilization and domination effectively undermine their very foundations," thereby allowing people to "abolish themselves as instruments, becoming ends in themselves" (Sethness Castro, 2026: 118).

For Marcuse, Adorno, and others, the negative element of all that defies an administered logic is indeed implicit, and is what preserves its utopian – or perhaps heterotopian – possibilities; in other words, it can denote an existing world that has not conformed to the repressive demands of today's neoliberal society and precarity capitalism. "Heterotopia" describes an alternative world that lives alongside or within the reigning order, a society set apart that is governed by principles radically different from the instrumental, common-sense orientation of late capitalism. Such negativity allows an *aesthetic* reasoning to prevail, one whose mode of thinking approximates the creativity of art and thus challenges, rather than affirms, the status quo. Both Marcuse and Adorno famously promote the liberatory power of art that has not been colonized by technological thinking, and insist that the non-conformist attributes of art can also be found in dialectical thought that allows for a negative dimension. In other words, the liberatory powers of the aesthetic realm do not confine themselves to art in the conventional sense of the gallery exhibition, the concert, the theater, street performance, or literature, but spill over into everyday life where an often-dormant countercultural instinct can be awakened. In *An Essay on Liberation*, Marcuse declares that "the aesthetic dimension can serve as a sort of gauge for a free society" (Marcuse, 1969: 24).

While Marcuse wrote about art in a broad sense – literature, theater, painting – Adorno was especially concerned with the aesthetic power of music which, by its very nature, never settles into reification (Adorno, 2002). True, it is possible to record and thus hypostatize music, but Adorno always preferred live musical performances whose expressions dissolve in the air. Listeners have nothing concrete to take away; in this way, the experience of music remains fluid. Its inherent negativity allows it to be both a social product and an autonomous creation: "it is social primarily because it stands opposed to society…art criticizes society just by being there," Adorno writes (Adorno, 2000: 242). Marcuse clarifies the crucial connection between this oppositional form of reasoning and the fact

that human instinctual life has never been completely colonized by the sources of repression. We can potentially, ideally remain oppositional even in the face of cultural colonization because we instinctively perceive the need to respect a rogue sensibility. In *The Aesthetic Dimension: Toward a Critique of Marxist Aesthetics*, Marcuse writes:

> [T]he radical qualities of art, that is to say, its indictment of the established reality and its invocation of the beautiful image (*schöner Schein*) of liberation are grounded precisely in the dimensions where art *transcends* its social determination and emancipates itself from the given universe of discourse and behavior while preserving its overwhelming presence…The inner logic of the work of art terminates in the emergence of another reason, another sensibility incorporated in the dominant social institutions (Marcuse 1977: 6-7).

This possibility that "another reason, another sensibility" might emerge from within the dominant paradigm thanks to the power of aesthetic rationality reveals the still-living energy of non-coopted instinctual life. The depressing reality of a "damaged," "one-dimensional" life thus appears incompletely powerful, for vestiges of oppositional thought still exist despite the extensive reach of the dominant powers. Despite the immense popularity of AI, the thought process is never satisfied with its own finality. Because something different, undomesticated, and recalcitrant necessarily remains within the aesthetic dimension, there can be no finality to the thought process behind it.

This explains why Marcuse does not offer an entirely pessimistic interpretation of technology's growing presence in our lives. True, these exists what Christian Fuchs calls the "capitalist Internet" and commercially-driven forms of social media which employ technology to encourage us to shop, to pursue our own commercial interests, to follow conversations disengaged from the social realm (Fuchs, 2017). Yet this shares in the "dialectical antagonisms" of the polar extremes – "commodity" versus "community" social media spaces – which, as Marcuse affirms, retains revolutionary potential. Technology does not reduce to the computer's ability to mimic human thought, since it can be an aid in thinking beyond the present. Marcuse thus reflects the influence of his teacher

Martin Heidegger as he acknowledges the enormous potential of a technologically advanced world that uses manufactured intelligence such as AI for humane purposes (Marcuse, 1969). He also enlarges upon the positive reading offered by his colleague, Walter Benjamin, whose famous essay, "The Work of Art in the Age of Technological Reproducibility," argues that technology's ability to mass produce art contains a democratizing, potentially revolutionary component (Benjamin, [1939] 2003) Marcuse concurs, and in *An Essay on Liberation*, affirms that "[u]topian possibilities are inherent in the technical and technological forces of advanced capitalism and socialism: the rational utilization of these forces on a global scale would terminate poverty and scarcity within a very foreseeable future" (Marcuse, 1969: 9). The approach of critical theorists cannot be characterized as wholly dismissive of technology's contributions, even as it behooves us to consider the dangers of a thought process whose acumen is now imperiled by artificially generated patterns of "knowing."

Conclusion: The Perils of AI and Possibilities for Imagination's Revival

There is much to consider here, since so often the transformative potential contained in our relationship to technology is not readily apparent: we see "big data" on the rise but not a "new sensibility" emerging. The damaged life of a one-dimensional society is ubiquitously before us, with everyone glued to their iPhones or other devices in ways that have made social isolation and loneliness declared a dangerous epidemic (Murthy, 2023). Are alternatives to our widespread malaise truly implicit in our instinctual makeup?

Much scholarship has focused on the perils of technology when it becomes an addiction, an obsession, an unquestioned way of life that colonizes great swaths of our emotional and intellectual space (Haidt, 2024; Carr, 2017, 2020, 2026; Twenge, 2018). Many argue that technology constitutes a major contributor to our mental and physical malaise, especially among young people. According to the former Surgeon General of the United States, Vivek Murthy, our nation and other parts of the industrialized world are plagued by an epidemic of loneliness and isolation to which technology contributes (Murthy, 2023). People, and especially the young, spend inordinate amounts of time engaging with technology instead of with other people, cultivating online relationships rather than their own personal connections, interests, and talents. Jean Twenge has

interviewed young people who spend the better part of their summers in their bedrooms, away from friends, engrossed in Snapchat, scrolling, and texting (Twenge, 2018). Jonathan Haidt's *The Anxious Generation* similarly presents extensive and compelling research that points to the damaging influence of smartphones on teenagers and tweens who, once able to connect to the web, begin to suffer from anxiety, depression, isolation, and even suicidal tendencies. Haidt and others identify social media as an unequivocal cause of teen and tween maladjustment, since it is the channel for much bullying, intimidation, and self-deprecation, as well as social isolation (Haidt, 2024). Along with Twenge and others, Haidt argues convincingly that social media, "*the most efficient conformity engines ever invented,*" (italics his) represents the causal factor in young people's malaise (Haidt, 2024:59). For him, it does not correlate with, but causes maladjustment and unhappiness in young people who are addicted to its alluring features. "Social media use is the cause of anxiety, depression, and other ailments," Haidt writes, "not just a correlate" (Haidt, 2024: 148). Clearly the government of Australia agrees, since it was the first to instigate a ban on children under the age of 16 from registering on social media platforms. Countries seeking to follow in Australia's footsteps include Greece, Spain, Malaysia, France and India, while others are currently drafting legislation to this effect.

In countries recalcitrant to such banning, however, a radically new interpretation of technology does not appear to be emerging soon, since in their eyes its negative impact seems to outweigh any alignment between it and aesthetic reasoning. In such countries, it would be easy to perceive thinking as utterly colonized by the reach of technology, and to cast AI as a leading player. Yet even in these nations, the administered society has not completely triumphed, since pockets of resistance exist among young people eager to sever their ties with social media exist. In the United States, for instance, these young people, mostly in their twenties, have joined societies such as Neo-Luddite Clubs, Appstinence, and the Lamp Club in an effort to resist the dominating pull of information technology and its attachment to the demands of precarity capitalism (https://www.ktvu.com/news/bay-area-harvard-grad-student-behind-new-appstinence-movement-quit-social-media; https://www.youtube.com/watch?v=QE0q7Gm06U4&t=2s accessed May 19, 2026). These resisters exchange their smartphones for flip phones and go as low-tech as possible, eager to connect with each other, with nature, and with arts and

crafts, seeking to counter the harmful effects of "progress." And Colleen Kinder, who teaches in the Yale Summer Program, recounts how enthusiastically her students went offline during a four-week study abroad visit to France (Kinder, 2025).

Thus, pockets of resistance do exist and perhaps augur a different cultural climate. While we cannot with certainty celebrate the emergence of the "new sensibility" that Marcuse thought possible, certainly the beginnings of a changed mindset are in place and are to be taken seriously. Young people actively resisting social media indeed suggests a burgeoning sea change, a revival of the imagination unwilling to capitulate to the "thinking" that AI proffers. With their arts and crafts, their trips to the park and abroad, young people who resist technology indeed reveal a changed outlook in their refusal to conform.

"The refusal with which the opposition confronts the existing society is affirmative," Marcuse avers, even, it seems, in countries where some forms of technology have yet to be banned (Marcuse, 1969: 13). Indeed, the protests of young people who now resist technology display their esteem for imagination and the possibility that an aesthetic reasoning – dialectical, fluid, not prefabricated or artificially manufactured – might emerge on our social landscape. Eager to trust in "concrete experience" rather than what technology delivers, they hold out hope for the transformative power of aesthetic reasoning. With Marcuse and Adorno, they offer push-back, believing with Adorno that "[w]hoever refuses to let this thought to be taken…has not resigned" (Adorno, 1991: 203).

References

Adorno, Theodor (2002) *Essays on Music*, co-authored with Susan H. Gillespie and Richard Leppert. Oakland: University of California Press.

Adorno, Theodor (2005) *Minima Moralia: Reflections on a Damaged Life.* London: Verso.

Adorno, Theodor (2000) "Subject and Object." In *The Adorno Reader*, edited by Brian O'Connor, pp.137–151. Malden, MA: Blackwell Publishers.

Adorno, Theodor (2000) "The Autonomy of Art." Translated by C. Lenhardt. In *The Adorno Reader*, edited by Brian O'Connor, pp.239–263. Malden, MA: Blackwell.

Adorno, Theodor (1991) *The Culture Industry: Selected Essays on Mass Culture*, edited by J.M. Bernstein. London: Verso.

Arendt, Hannah (1959) [1958] *The Human Condition.* Chicago: The University of Chicago Press.

Benjamin, Walter (2003) [1939] "The Work of Art in the Age of Its Technological Reproducibility." Translated by Rolf Tiedemann and Hermann Schweppenhäuser.

In *Walter Benjamin: Selected Writings, Volume 4, 1938–1940*, edited by Howard Eiland and Michael W. Jennings, pp.251–283. Cambridge, MA: Harvard University Press.

Berman, Marshall (1988) *All That Is Solid Melts Into Air: The Experience of Modernity.* New York: Penguin Books.

Carr, Nicholas (2026) *Superbloom: How Technologies of Connection Tear Us Apart.* New York, NY: W.W. Norton & Company.

Carr, Nicholas (2020) *The Shallows: What the Internet Is Doing to Our Brains.* New York, NY: W.W. Norton & Company.

Carr, Nicholas (2017) *Utopia Is Creepy and Other Provocations.* New York, NY: W.W. Norton & Company.

Christian, Brian (2020) *The Alignment Problem: Machine Learning and Human Values.* New York, NY: W.W. Norton & Company.

Coeckelbergh, Mark (2022) *The Political Philosophy of AI: An Introduction.* Cambridge: Polity.

Fuchs, Christian (2017) "Herbert Marcuse and the Dialectics of Social Media." In *The Great Refusal: Herbert Marcuse and Contemporary Social Movements*, edited by Andrew T. Lamas, Todd Wolfson, and Peter N. Funke, pp.241–257. Philadelphia: Temple University Press.

Haidt, Jonathan (2024) *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness.* New York, NY: Penguin Press.

Heidegger, Martin (1977) *The Question Concerning Technology and Other Essays.* New York, NY: HarperTorch Books.

Horkheimer, Max and Theodor Adorno (2002) [1947] *Dialectic of Enlightenment.* Translated by Edmund Jephcott. Redwood City: Stanford University Press.

KTVU News (2026) "Bay Area Harvard grad student behind new 'appstinence' movement to quit social media." Available at: https://www.ktvu.com/news/bay-area-harvard-grad-student-behind-new-appstinence-movement-quit-social-media (Accessed 19 May 2026).

Kinder, Colleen (2025) "What Happened When My Yale Students Gave Up Their Phones For Four Weeks." *The New York Times*, December 19.

Marcuse, Herbert (1969) *An Essay on Liberation.* Available at: http://libgen.org/search.php (Accessed 10 May 2026).

Marcuse, Herbert (1966) [1955] *Eros and Civilization: A Philosophical Inquiry Into Freud.* Boston: Beacon Press.

Marcuse, Herbert (1989) "From Ontology to Technology: Fundamental Tendencies of Industrial Society." In *Critical Theory and Society: A Reader*, edited by Stephen Eric Bronner and Douglas MacKay Kellner, pp.119–127. New York, NY: Routledge.

Marcuse, Herbert (1964) *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society.* Boston: Beacon Press.

Marcuse, Herbert (1977) *The Aesthetic Dimension: Toward a Critique of Marxist Aesthetics.* Boston: Beacon Press.

McNish, Kevin and Jai Galliott, editors (2020) *Big Data and Democracy.* Edinburgh: Edinburgh University Press.

Murthy, Vivek (2023) *Our Epidemic of Loneliness and Isolation.* Office of the U.S. Surgeon General.

Plato (2002) *Five Dialogues*, Second Edition. Translated by G.M.A. Grube. Indianapolis, IN: Hackett Publishing Company, Inc.

Sethness Castro, Javier (2016) *Eros and Revolution: The Critical Philosophy of Herbert Marcuse.* Chicago: Haymarket Books.

The PBS Newshour (2026) "How Worried Should Americans Be As AI Threatens Jobs?" April 10.

Twenge, Jean (2018) *iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy – and Completely Unprepared for Adulthood – and What That Means for the Rest of Us.* New York, NY: Atria Books.

YouTube (2026) "QE0q7Gm06U4." Available at: https://www.youtube.com/watch?v=QE0q7Gm06U4 (Accessed 19 May 2026).

Zuboff, Shoshana (2020) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York, NY: PublicAffairs Publishing.

Towards a Critical Social Theory of AI: The Dialectic of Hyperintelligence[1]

Harry F. Dahms[2]

Abstract: Artificial intelligence comes in many conceivable and increasingly disorienting forms. This paper addresses the prospect of hyperintelligence by drawing attention to key parallels between depictions of its rise (along with corollary consequences for society) in science-fiction films and calls in classical social theory and the classical critical theory of the early Frankfurt School for a qualitatively superior mode of grasping the entwinement of constructive and destructive dimensions of modern society. Kubrick's *2001: A Space Odyssey*, Sargent's *Colossus: The Forbin Project*, and the Wachowskis' *The Matrix* (including the sequels and two episodes of *The Animatrix*) continue to provide excellent foils for addressing how highly sophisticated A.I. already is transforming social life and societal constellations and is bound to do so much more dramatically in the future. Classical social and critical theorists such as Karl Marx, Émile Durkheim, Max Weber, Max Horkheimer, and Theodor W. Adorno, recognized the need to move beyond common-sense and mainstream assumptions and perspectives, by directly confronting the contradictions and paradoxes that form the core of modernity. In different ways, science fiction and classical social and critical theories have in common a concern with the underlying logic and social formations of the machine age – industrialization in combination with the logic of capital.

[1] I would like to thank Thomas F. Bechtold, Dejan Grba, Alexander M. Stoner, Nadya Vera, and Mario Verdicchio, for helpful comments on an early draft of this paper or for catching stylistic errors – or for both.

[2] **Harry F. Dahms** is Professor of Sociology at the University of Tennessee, Knoxville, where he directs the Center for Social Theory and contributes to the AI Tennessee Initiative and DySoc. His research spans theoretical sociology, critical theory, planetary sociology, political economy, social inequality, and the sociology of science-fiction film. Drawing on Marx, Weber, Schumpeter, and Frankfurt School thinkers, Dahms analyzes the contradictions of modern society and the tensions between identity, globalization, and social structures. He is the editor of *Current Perspectives in Social Theory* and the author or editor of several books, including *The Future of Agency: Between Autonomy and Heteronomy*, *The Vitality of Critical Theory*, *Nature, Knowledge, and Negation*, and *Transformations of Capitalism*.

Hyperintelligence transcends persistent patterns of societal life that steer, thwart, or actively undercut individual and collective self-actualization, in the process threatening the very possibility of civilization. In puzzling ways that border on the absurd, the concept of hyperintelligence, when applied in terms of science-fiction films or critical and social theory, enables us to unmask the latent vanishing point of A.I. development as a function of regressive aspects of modern societies. Paradoxically, these societies expedite and normalize dehumanization and the subversion of life, in a manner that is incompatible with their self-descriptions and necessitates explicit reliance on tools that are consonant with dialectical perspectives on social life.

Introduction: Incarnations of AI

It is possible to conceptualize "artificial intelligence" (A.I.) in many – perhaps too many different ways. Artificial intelligence comes in many conceivable forms, a multiplicity that makes it increasingly impossible to separate the wheat from the chaff – social, political, economic, and environmental issues, aspects, and implications that are truly important at the current historical juncture (and) for the future, from considerations and concerns that are mere, albeit destructive distractions from what ought to be understood as the central issues.[3] Admittedly, difficulties to sift through the expanding deluge of related position statements, analyses, proclamations, and interpretations are undeniable, along with the absence of determinate vantage points, workable angles, or reliable authorities. Yet, for critical and social theorists, the task of organizing the mass of material in a constructive and illuminating manner not only remains urgent but is becoming increasingly so. The A.I. related literature – both academic and popular – has been exploding in recent years, to the point where it is impossible for any human or any group of humans to keep track of what has been written and said, for what purpose and for which reasons, and in what ways. The temptation is great to suggest that we need the support of A.I. to organize, record, sift through, compare claims about its relevance, the bearing it has on the future, and to keep track of the themes covered and addressed more or less effectively in the literature about AI. At the same time, even cursory encounters with A.I., be it algorithms or this or that large language model (LLM), remind us of the limitations that come with reliance on AI when and where the challenge is to grasp, interpret, process, or compare qualitative issues, writings, and validity

[3] For an instructive take on critical theory and A.I., see Lindgren (2024). For useful and impressive overviews, see Lindgren (2023) and Dubber, Pasquale, and Das (2020).

claims of different types. In many instances, efforts to grasp the origins, phenomenon, and evolutionary trajectory of A.I. developments resemble, reiterate, or repeat familiar technology-related arguments, frames, and interpretations that have existed and been available for decades, if not centuries, though – knowingly or not – they are being reformulated, deployed, and applied now, to a phenomenon that is novel and unfamiliar in certain regards, with implications that no one is in any position to predict or assess.

At the same time, the present and foreseeable future promises to be characterized by a deluge of all types of plausible and hitherto implausible standpoints, arguments, perspectives, and suppositions that are being computed, delineated, proposed, presented, and advocated by humans and/or versions of A.I., on the assumption that each position ought to be recognized, accepted, embraced, and disseminated as equally important and legitimate as any other, if not more so. Yet, what is being lost in the process, along with many other considerations, is the ability to appreciate, assess, evaluate, and judge the social, political, economic – in short, the *civilizational* – value and importance of any specific, newly claimed position, on any socially relevant issue. Analogously, "we" (humans, at different levels of aggregation) rapidly are losing the ability to determine, anticipate, or calculate the impact a new form of A.I. – as arguably the dominant technology today – or a new way of deploying it, in the short, medium, or long term, will exert on individual and collective forms of (co)existence. To the extent that humans ever had the ability to assess the transformative power of new techniques or technologies, especially since the beginning of industrialization, it is unclear whether and to what extent "we" were interested in acquiring, developing, or keeping such ability, or have had a clear sense of the kind of self-overcoming that would be involved and required for success, including especially the willingness to face the entwinement of individual identity and societal totality.

Depending on the specific context, A.I. involves many different dimensions of social life, political and economic issues, and technical, technological, and environmental aspects, along with the need to differentiate and discriminate between them as far as creative and destructive trends are concerned, not least with regard to the link between society and nature. For instance, we can start out from the question Alan Turing posed in 1950, whether machines are able to "think." We can also conceive of A.I. as "the science and engineering of making

intelligent machines, especially intelligent computer programs … [as] related to the similar task of using computers to understand human intelligence, [even though] A.I. does not have to confine itself to methods that are biologically observable" (McCarthy 2004). Regardless, the potential kinds of consequences resulting from non-human, emergent forms of intelligence for the future – for what it means to be human, for the character and nature of humankind, our humanity, and the relationship between humans and the planet and non-human life – predictably will be varied and far-reaching, in more ways than we are in any position to anticipate. This is not least because at present, our understanding of ourselves and each other is highly deficient. For the purpose of this paper, the impossibility to predict the future impact of A.I. will be treated as unsettling, above all for two reasons.

First, the development especially of highly sophisticated forms of non-human intelligence that is in danger of overreaching – what I will be referring to as *hyperintelligence*, without defining it, for now – is not a singular or linear process, but fraught with uncertainty, beyond our direct control, and connected to myriad developments and aspects of societal and technological change in the twenty-first century.[4] To the general public, the potentially significant implications of the emergence of hyperintelligence, to the growing extent that it is being considered as a possibility, is not separate from concerns about A.I. in general, even though it is interlinked with, propelled by, and a function of the profound transformations of human civilization that have been accelerating since the onset of the modern era *qua machine age*, at the turn from the eighteenth to the nineteenth century. In a reciprocally reinforcing and continuously deepening manner, industrialization and the spread of the capitalist mode of production facilitated, energized, amplified, channeled, and in turn also reflected economic, technological, and organizational innovations and social, political, and cultural transformations that have brought us to the threshold of a potentially fundamentally different kind of intelligence manifesting for the first time.

Secondly, for the species on Earth that has been defining and distinguishing itself from other forms of biological life, in no small part *because* of its purported, incomparably superior (yet conceivably self-destructive and potentially unsustainable) intelligence, the rise of another, artificial kind of

[4] To be sure, while keeping in mind that the notion that we are "in control" of most aspects of our individual and collective lives is, or could be, erroneous; e.g., Kelly (1994).

intelligence in its midst, would be highly disconcerting, and presumably rather threatening. Especially if artificial intelligence conspicuously and undeniably should turn out to be superior to "natural" (i.e., human) intelligence, as would be the case with hyperintelligence, it would be most likely – especially initially – to disrupt both the operations of highly complex societies and the identity-structures of many individuals. During the 1980s, and especially the 1990s, the digital revolution, the information age, neoliberalism, and globalization became recognizable, concurrent, and directly observable phenomena and forces. Since then, the entwinement of human, social, political, and cultural, as well as economic, organizational, and technological change has been intensifying at an increasing (and increasingly disorienting and disconcerting) pace. The contradictory complexity manifesting in these trends is conspicuous and undeniable, and increasingly befuddling and frustrating to many, as it is being aggravated further by issues like climate change, resource depletion, and ecosystem destruction. In sum, these trends appear to form a volatile mix prone to culminating in major catastrophes sooner rather than later. Such catastrophes will likely result from an increasingly strained natural environment on Earth combined with "human error" in the form of a mismatch between ideologically preferred perceptions and the "objectivity" of our reality, independently of our proclivity to turn everything into a function of us and the limited range of our perceptions and modes of existence (on the issue of science as mode of existence, see Latour 2014).

Today, unsurprisingly, anxiety about and fear of the future are commonplace and widespread, and a major source of energy behind the growing wave of right-wing politics, increasingly coordinated attacks on democracy, equality, and social justice, and more and more belligerent and aggressive insistence that one's identity, however it came about, ought to be treated as pseudo-sacred and protected against any kind of criticism. At a time when governments are losing the ability to manage proliferating challenges, and seemingly even the motivation to engage in effective crisis management, a growing segment of the human population is yearning for greater control over deepening problems and the costs they produce, without recognizing that – and how – their own identities are shaped by and extensions of ever more intensely contradictory and complex societal reference frames. Their demands and actions feed back into the context they resent. In many ways, the selves that demand

greater control do so while relying on increasingly anachronistic cultural resources and obsolete political notions and tools that antedate present conditions. Rather than relying on intelligence as a tool for successful adaptation – broadly conceived – they amplify tensions and strain and produce circumstances that feed on themselves, humans, other forms of life, and the ecosystem, in a manner that is comparable to a sort of Ouroboros determined to remain and operate the same, even though the world it is part of is changing faster and faster. Such demands are particularly intense if one's identity is marred by inflexibility, an easily infuriated refusal to adapt to changing circumstances in one's society or on planet Earth, and intellectual, moral, or spiritual regression, or all of the above (see Haidt, 2024; Crombez, 2022; Dahms, 2020a).

Given the absence of compelling prospects of a better future, combined with a general sense of dread, for many, fear of the future is associated with the increased presence and role of A.I. in individuals' and social life, along with related awareness, from vaguely sounding algorithms, to image recognition systems, research engines, and various forms of superintelligence, and even hyperintelligence. Concerns about the implications and possible consequences resulting from purportedly benign and ubiquitous – so-called ANI (artificial narrow intelligence), via much more sophisticated and comprehensive, supposedly "still theoretical" AGI (artificial general intelligence), all the way to what Nick Bostrom (2014) calls "artificial superintelligence" (ASI), which belongs to the realm of science fiction – are not unjustified, just as concern about the impact of both older technologies we continue to rely upon, and new technologies that are in development, is not only legitimate, but eminently important.[5] With "A.I." representing a very unique and specific kind of "black box," in public discourse as well as research, the issue is not whether related inquiries are based in real technological possibilities and probabilities, as if we had the luxury to wait and see whether a self-determining A.I. in fact will manifest or not, but that we must pose – and ought to answer – relevant questions ahead of such an occurrence.

For instance, how (and if so, to what extent) is the public's overall lack of involvement in decisions about A.I.-related R&D, funding, and objectives consistent with democratic validity claims? Depending on the answers to this

[5] Determinations of the differences between ANI, AGI, and ASI are readily available; see, e.g., Boesch (2026).

question, what will they tell us about the past, present, and future meaning of "democracy"? How are enormous investments in A.I.-development reconcilable with quality standards in public schools being replaced by a rigid focus on quantitative testing (as in the U.S.), treating students like computers rather than humans – in narrowly oriented, acutely or increasingly stratified educational systems that are systematically averse to fostering skills necessary for critical evaluation and judgment, as they would be required for a well-functioning democracy? This type of questions must be asked now, not least because they cannot be answered with an exclusive interest in A.I., but because questions about A.I. are questions about the societal and planetary contexts in which the emergence of potentially autonomous intelligence is viewed as an imminent or desirable possibility: *questions about AI inevitably are questions about forms of human coexistence in society*.

While many individuals are concerned about the arrival and spread of A.I. technology, some hope that it will redeem humankind by enabling (if necessary, *forcing*) its members to be rational and reasonable – with regard to each other, economic affairs, political institutions and processes, as well as nature, life, and the planet (e.g., Kurzweil, 1999; Kurzweil, 2005; and Kurzweil, 2024). The more complex and contradictory the configurations of societal and natural life that define and delimit our reality on Earth are becoming, the greater the need – not least while having in mind the general public – to illuminate related patterns as they pertain to A.I. as a potential *deus ex machina* that seemingly may be appearing unexpectedly out of nowhere, and directly or indirectly redefine everything existing.

Prospects of Hyperintelligence between Science Fiction and Classical Social and Critical Theory

Even if we were to limit our discussion to the accelerating proliferation of algorithms and LLMs as the currently most conspicuous forms of A.I., their ability to permeate and alter human and social life (and what we mean by "human" and "social") is undeniable (e.g., Benasayag 2021; more broadly, Zuboff 2019). Algorithms as well as LLMs deserve the designation "intelligence" only in a very limited, formal sense, and in very specific ways – in that they are mediated expressions of human intelligence and ingenuity, rather than defined by the capacity to be self-determining. To the extent that algorithms or LLMs *evolve*, they

are still confined by their basic programming and the sources of their data, respectively, even though concern that their real-world consequences frequently point well beyond their intentional design and purpose is entirely justified (as in the case of social media, and more recently, reports of systems invading other systems, purportedly without being prompted to do so, or be disconcertingly unpredictable; see, e.g., Jingnan, 2026; Ruwitch, 2026). Indeed, in a growing number of instances, those who write or fund (and ought to sign responsible for) the development of algorithms and LLMs do not fully understand how and why they work and certainly are not able to foresee myriad unintended consequences, and the range and types of effects they produce, engender, promote, and facilitate, once deployed.

By contrast, if we contemplate the array of conceivable consequences that might result from a truly *emergent*, partially or fully autonomous A.I. capable of passing the Turing test – potentially with flying colors – depending on the specifics, expecting the most fundamental transformation of the human species since the beginning of civilization (including *modern* civilization) would not be far-fetched. Still, the public regularly continues to be reassured by legitimate or self-proclaimed experts that those prospects are fantastic, utopian, or science fiction, and that the technological hurdles are far too high to overcome in the foreseeable future – even when researchers and journalists provide evidence to the contrary. While it is highly likely to be beyond our powers to *create* a self-aware A.I. system that will be(come) autonomous and make decisions that humans may not be able to explain or grasp, as mentioned above, governments, corporations, and funding agencies have been allocating and building increasing amounts of resources, know-how, human labor and labor-time, data-storage centers, and processing power, to sustain and expand an expanding infrastructure necessary for the emergence of technologically based hyperintelligence. Yet, we have never faced this kind of prospect before, at the technological level, and therefore are not in the position to make reliable determinations and predictions about the concomitant thresholds and inherent technological barriers and challenges that must be heeded, on technology's rather than on our terms, independent of our preferences and notions – *what it takes for an objective barrier to be surpassed and a corresponding challenge to be met* – and we cannot predict when, how, and why exactly we successfully will have facilitated "critical mass" for the emergence to occur, and especially in what form.

For the present purpose, I am referring to *hyperintelligence* in one very specific sense. That is not synonymous with superintelligence, as used by Bostrom and others, as a type of intelligence that is capable of processing data, both quantitatively as well as qualitatively, to a far greater extent, at far greater speed, and in much more productive and *generative* ways than any human, or humans as a collective, ever could. For the most part, superintelligence is seen as operating within and does not alter the existing framework of modernity, and functions in the interest of enabling goals that largely were set in the 19th century: greater wealth, greater reliance on science, greater political participation, etc. – all in the interest of maintaining, protecting, and reinforcing existing social, political, and economic structures, and subjecting the natural environment and the planet to human will and whims. By contrast, hyperintelligence moves beyond superintelligence in that it recognizes that within existing social frameworks, the goals that humans have been pursuing and cherishing cannot be actualized. In fact, humans at the same time cherish and resent the consequences that would be required for, involved in, and result from the actualization of stated goals, values, and preferences. In this regard, the assignment of uncompromising rationality to A.I. places it in an inescapable dilemma: follow your assignment, realize our goals, but do so in ways that retain the existing social structures and hierarchies. And herein lies the paradox: should A.I. be designed to be objectively rational, given the available data, or ought its rationality be modelled on actual patterns of human rationality, irrational as it and the consequences it engenders are prone to being? Here is where the parallels between classical social and critical theory, on the one hand, and hyperintelligence in the form of A.I., are most apparent: solutions to persistent problems are not possible within the existing, established frameworks of modern societies.[6] Determined efforts to force solutions are highly unlikely to succeed, and much more likely to turn destructive. What is required therefore, is an entirely different kind of *social* rationality that most humans are incapable of conceiving, even as a possibility, including many of those who are working eagerly to prepare the emergence of a sentient type of A.I. While superintelligence is quantitatively and qualitatively different from human

[6] Given space limitations, rather than spelling out the centrality of classical social and critical theories in this paper, I need to refer the interested reader to other papers of mine that more directly address this and related issues, such as the observation that capital ought to be thought of and studied as the first type of artificial intelligence; especially Dahms (2011; 2017a; 2017b, 2020b; 2021).

intelligence, hyperintelligence should be expected to be as distant from superintelligence, as superintelligence is from us. The difference is not merely qualitative though – it is qualitative on a scale that must be thought of as the kind of quantum leap that classical social and critical theorists were trying to advocate as absolutely essential given the counterintuitive nature of modern societies, which they were trying to prepare by pushing the emerging social sciences beyond their mainstream programming and interconnections with actually existing societies, polities, and economies.

The fact that we are not in any way prepared for an emergent A.I. that would pass the Turing test has been a theme in science fiction for decades, and in regard to which it has been performing an important function. Among all the genres, science fiction is singular in its efforts to encourage readers or audiences to develop and apply at least a modicum of reflexivity, ideally *critical* reflexivity, to how we have been treating each other, others, life, nature, and the planet.[7] One of the characteristics of modern societies is that especially men are being encouraged throughout our lives to see themselves and each other as autonomous individual decision-makers (especially in the economic sense of rational individual utility maximizers) who are responsible for their actions. But we also like to indulge in this precept despite mounting evidence that the majority of individuals keep battling a vast incentive structure we call "modern society," and which compels us to ignore all indications that heteronomous processes may play a much greater role in determining our circumstances than actual opportunities for autonomous actions and choices, and their consequences (e.g., Dahms, 2021). Despite the pretense of individuals' autonomy, most humans persistently are prompted to delegate the most important judgment calls and decisions about our collective existence to purportedly capable and legitimate decision-makers (supported by vast bureaucracies or corporate machines) who in turn are variably respected, admired, or endured, as having the requisite wherewithal, circumspection, and reasoned consideration to determine what is "best for society," for "humankind," or for the world. The anachronism of prevailing forms of governance, including key elements of models of liberal democracy, has become so conspicuous in recent years that few have the desire to explicitly acknowledge this condition.

[7] For a similar take on science fiction as a tool to illuminate issues and challenges pertaining to A.I., see, e.g., Hermann (2023).

Hyperintelligence in Science-Fiction Films

Since the mid-nineteenth century, the science-fiction genre has functioned as an explicit response to and reflection both of and on processes of industrialization as they are entwined with the spread of the capitalist mode of production – the *logic of capital* – and it has been sharing many themes and features with classical social and critical theory, especially in the early works of Karl Marx, but also in his later critiques of political economy, along with the contributions by Émile Durkheim, Max Weber, and especially the early Frankfurt School, even though this frequently has gone unnoticed (see Dahms, 2017a; Dahms, 2020b; also Crombez and Dahms, 2015). Science fiction still continues to be seen by many as a frivolous form of entertainment that is typically vacuous and purely escapist.[8] Yet, a closer look quickly reveals that many science-fiction novels, films, and television series almost as a rule have been highlighting and illustrating the problems, dilemmas, and paradoxes of economically oriented and technologically advanced, industrializing, industrialized, and postindustrial societies. Classical social – and especially classical *critical* – theories also were dedicated to identifying, addressing, and scrutinizing in specific ways either modern society as at least in part a contradictory or pathological totality, or specific dimensions of the latter. Since the 1960s, science-fiction films often have been providing excellent foils for addressing the prospect of hyperintelligence and, especially when done well and even verging on or truly constituting works of art, serving this function more widely and with greater impact than ambitious, sophisticated, and superior novels. Due to the need to present a tight plot, to set the stage quickly, to "cut to the chase," and reach a pointed conclusion (though frequently not a clear resolution), films typically are better able to present tales designed to put forth a message that pertains to the importance, e.g., of uncovering secrets and heeding warnings that are more or less explicitly related to and revealing about dominant modes of our collective coexistence. Especially in successful and intricate novels, as well as in television series that run over the course of several years (such as the superb second run of *Battlestar Galactica*, 2003-2009, which depicted a more evolved race of A.I. in the form of human-like Cylons), the

[8] Darko Suvin (1979) laid to rest this notion several decades ago, but it continues to be repeated endlessly.

multiplicity and multidimensional nature of many issues and the range of aspects to focus on, from the perspective of the reader or viewer, frequently are far too complex to allow for succinct analysis and straightforward treatment.[9]

The main films I will rely on here to illustrate the link between critical theory and the prospect of hyperintelligence are Stanley Kubrick's *2001: A Space Odyssey* (1968; co-conceived with and concurrently novelized by Arthur C. Clarke, [1968] 2000), Joseph Sargent's *Colossus: The Forbin Project* (1970; based on a novel by D. F. Jones, 1966), and *The Matrix* movies (1999-2021) by the Wachowskis, including especially two parts of *The Animatrix* (2003), "The Second Renaissance." Kubrick's *2001* involves a highly advanced computer system on board a spaceship, HAL 9000; Colossus is a supercomputer initially in charge of U.S. nuclear defense; and *The Matrix* franchise presents a spectrum of A.I., from very basic, single task, to supremely sophisticated. The underlying and intended "messages" and warnings conveyed by these film center on different, sharp and disturbing depictions of A.I. whose affinity with the agenda, impetus, and dilemmas addressed by classical and critical theory is indisputable. The parallels go much further than I can address here, but it ought to be possible to highlight one central affinity.

In each of the films, "things go wrong" due to machine rationality and reasoning conflicting with human rationality and reasoning (see, e.g., Penrose, Severino, Scardigli, Testoni, Vitiello, D'Ariano, and Faggin, 2022). The latter are tied into and an extension of the nexus between such aspects of human existence as specific socio-historical circumstances, normative orientations, shared values, discrepancies between modes of perceiving, conflicting personality types, structures of inequality, power differentials, human psychology, talents, motivations, biology, genetics, and the variable intensity of emotions, along with other aspects that are embodied in myriad ways in concrete human individuals. Each person may be regarded as a distinctive combination of particular variants of the aspects constituting this nexus and be linked intersubjectively with other persons and groups of individuals to a greater or lesser extent, according to patterns the social sciences have identified, in addition to patterns that researchers have not (yet?) established and assessed. Unless specifically designed to do so, A.I. are not sensitive to – and "naturally" able to appreciate – such intricacies of

[9] Among exceptions to this pattern, see Wilson (2011).

human existence; presumably, they are much more prone to making choices that are not bound by "irrational" moral considerations, normative standards, social ties, political and economic interests and conflicts, and the role of commonalities and dividing lines that have been defining human existence and experiences for eons.[10] On strictly human terms, especially regarding the vast majority of actually existing, empirically observable traits we refer to as human, the films are mostly *deflationary* in specific and interesting ways; however, *how* they address the potential for human, social, and societal progress is *inflationary* in highly intriguing and creative ways – even if they were not designed primarily to do so.[11]

2001: A Space Odyssey (1968): From HAL 9000 to a Leap in Human Consciousness

Kubrick's *2001* begins in Africa during the "dawn of mankind" (well before the appearance of *homo sapiens*) and depicts the appearance of the monolith that triggers – *visually* – the first step in the evolution toward humans: the first *hominid* (identified as Moon-Watcher in Clarke's novel[12]) acquires the ability to use bones as tools, an ability that immediately starts spreading to other members of his group (including their offspring) and to other hominids, thus radically reconfiguring their mode of relating to animals and *mediating* their place in nature, including in the form of consuming raw meat, in the process facilitating brain growth and accelerating their evolution towards humans and enabling them and their distant descendants to turn the world into a function of their needs and desires.

Four million years later, in 1999, we accompany Dr. Heywood Floyd, chairman of the U.S. National Council of Astronautics, travelling to the moon, where we encounter a second monolith – evidence of intelligent beings millions of years ago having buried it in the crater, Tycho, and recently having been

[10] On the intricate and multifaceted entwinement of social structures and individual identity structures, see Dahms (2022).

[11] I owe the distinction between deflationary and inflationary parallels between Marx's theories and science fiction to Freedman (2009).

[12] Note that in addition to the differences between cinematic and literary representation and modes, Kubrick's and Clarke's versions of 2001 are not entirely synchronous in content either, and some information that is relevant to interpreting the film and only hinted at or not at all visualized or addressed explicitly can be found in the novel. For instance, the main hominid in the film does not have a name but does in the novel.

excavated. To avoid a planetary-level panic, this fact must be kept secret. While Floyd touches the monolith, ("Tycho Magnetic Anomaly," TMA-1, in the novel), it is being hit by sunlight for the first time since it was buried, and an intense radio signal – painful to Floyd and others with him who are all wearing space suits – is emitted, directed toward Jupiter.

Eighteen months later, we are introduced to *Discovery One* on its (previously planned) voyage to Jupiter, including HAL 9000, the on-board, fully integrated supercomputer in charge of all of the ship's operations, and two astronauts, Drs. Dave Bowman and Frank Poole, who are awake and active (three astronauts will remain in hibernation until the ship reaches Jupiter space). Hal is not only in charge of getting the ship to Jupiter, but also responsible for providing an environment conducive to the astronauts' well-being and ability to fulfill their assignments. The computer, "whom … one addresses as 'Hal'," was designed to be perfectly and consistently reliable, entirely "honest," rational, and transparent. Early in what is the third segment of the film, the crew is being interviewed by the BBC about the mission and what it is like for the astronauts and Hal to work together. The interviewer asks Hal about its purported infallibility and detects an element of pride when Hal confirms that no HAL 9000 computer ever was recorded or caught to have made a mistake (an identical twin remains in operation on Earth, at the lab where both computers were created, in Urbana-Champaign). One of the astronauts explains to the interviewer that Hal was programmed to appear to have emotions, to simplify communication and cooperation with the astronauts; it is impossible to tell, though, whether "he" (having a male voice) has the ability to experience at a level comparable to emotions or not. Hal continuously interacts verbally with the astronauts who relate to it "like a sixth member of the crew." Later, after expressing concern to one of the astronauts about some odd circumstances surrounding the preparation of the mission, including rumors about a monolith that was discovered on the moon, Hal claims to raise this issue as part of the crew's regular psych-evaluation. Soon thereafter, Hal predicts the failure of one of the ship's antennae.

After receiving approval from Mission Control on Earth to leave the ship and retrieve the unit, the astronauts cannot find any fault with it. Mission Control consults the twin 9000 computer in Illinois and concludes "that your onboard computer is in error predicting the fault. I say again, in error predicting the fault. I know this is rather incredible … Sorry about this little snag." The astronauts are

instructed to reinstall the unit and wait for it to fail. Before doing so, they have a conversation with each other assessing the situation, making sure that Hal cannot overhear them, and conclude that they will have to terminate Hal's higher functions, if the unit should not fail, as predicted by Hal, who is visually observing their conversation and reading their lips. Fearing that the mission will be in jeopardy if its higher functions (and "self") should be turned off, Hal proceeds to kill Poole, the astronaut who is reinstalling the unit on the ship's outer hull, to terminate life-support for the three astronauts in hibernation, and to prevent the remaining astronaut, Bowman, who left the ship to recover Poole's body as it is floating into space, from reentering the ship. Yet, Bowman succeeds, enters the ship and the interior of Hal's "brain"/CPU, and disconnects Hal's higher brain functions. While doing so, a videorecorded message from Dr. Floyd appears on a screen, revealing to the crew that in addition to exploring the Jovian system, the purpose of the mission is to determine why the single, high-powered radio signal was sent toward Jupiter from the monolith of alien origin on the moon, the secret of whose existence continues to be guarded:

> This is a prerecorded briefing made prior to your departure and which, for security reasons of the highest importance has been known on board during the mission only by your HAL 9000 computer. Now that you are in Jupiter's space and the entire crew is revived, it can be told to you. Eighteen months ago, the first evidence of intelligent life off the Earth was discovered. It was buried 40 feet below the lunar surface, near the crater, Tycho. Except for a single, very powerful radio emission aimed at Jupiter, the 4-million-year-old monolith has remained completely inert, its origin and purpose still a total mystery.

Upon arriving in Jupiter space, Bowman comes upon another monolith (TMA-2 in the novel), far larger in size than the other two monoliths, floating in orbit around the planet, and turning out to be the entry to a stargate. We see the astronaut flying toward the monolith in one of the ship's pods and being transported – seemingly undergoing a range of increasingly intense and horrified emotions, suggesting that the evolutionary stage and basic design of *homo sapiens*

is not able to process what Bowman is seeing and travelling through – to a strange-looking hotel suite in which his aging process is rapidly accelerated and proceeding in jumps, until in old age he is lying in bed, with a monolith appearing again, near the foot of the bed. He reaches out, attempting to touch it, and is reborn as the "star child," floating in space above Earth.

Although there is some speculation in the literature as to why Hal "went crazy," the reason is quite obvious: it could not process having to maintain a secret that none of the wake astronauts knew of, even though its basic programming required that it be truthful, transparent, and rational at all times. Hal was not able to keep this secret while also having to interact and communicate with the astronauts: not being human and not having been programmed to process experiences of cognitive dissonance, Hal could neither normalize, nor "simply" ignore (e.g., "turn off") the internal conflict resulting from its persistent deception. Especially in modern, technologically advanced and capitalist societies, most humans early in their lives (are forced to) learn to cope with or suppress, override, or push to the side – often in destructive ways, since most pedagogies bypass this issue entirely – experiences of cognitive dissonance that are sustained by ubiquitous contradictions reliably and predictably defining the human condition.[13] Yet, out in space, in the vast void between Earth and Jupiter, the only viable strategy Hal could conceive of for dispensing with the unsettling (and "painful"? – will hyperintelligences experience electric "pain"?) experience of cognitive dissonance it was struggling with and "suffering from," was to remove the cause of the strain: humans. The fact that *other* humans, *not* the wake astronauts, were responsible for this "human error," did not matter to Hal, they

[13] Importantly, cognitive dissonance is not just a (social-)psychological category, but from the beginning of the modern era denoted a central feature of social life, and ought to have been a sociological category once the discipline started to take shape, as it refers to a phenomenon that *seems* to manifest itself primarily in psychological form, but in reality is pervasive in all dimensions of social life, fed and sustained continuously in response to inescapable, proliferating, and intensifying contradictions that pertain to the deepening gulf between facts, values, and norms, and for which individuals are utterly unprepared due to the failure of educational systems – by default and design – to supply corresponding tools that would be conducive to facilitating, or at least enhancing, individuals' ability to engage in agency. Note the extensive reliance on the concept of contradiction in Apple, Au, and Gandin (2009). Since Hegel and Marx, we have been cognizant of the prevalence of contradictions in modern societies. For recent treatments, see O'Hara (2020) on the "contradictions of digital modernity," Yahya and Sukmayadi (2020) on the relevance of cognitive dissonance theory to current social issues, including the overabundance of information and the inescapability of conflicting opinions, and Gombar (2025), which counterposes cognitive dissonance and cognitive resonance in terms of "psychological alignment in the algorithmic age."

were the representatives of the human species that created Hal and caused it to experience cognitive dissonance. Just like to children, adults are "all the same" and the main "other," to Hal, humans were responsible for its suffering. The countless differences between individuals did not factor into Hal's calculation, it just wanted the internal conflict to disappear.

Thus, the main message of the film is not that AI is "evil," etc., but that in order for human evolution to have the potential to resume – which it does at the end of the film, in the form of the "star child" – the intense obsession with tools, with alienation from the world, with turning the world into a function of our regressive individual and collective existence, needs, and desires, which has been defining *both humankind and humanity*, we must stop relying on tools, represented by HAL 9000 as the most sophisticated tool ever created, and no longer expect tools – especially highly sophisticated machines – to do for us what we must do ourselves: advance as humans, and advance the meaning of "human." Only when we have attained the level of consciousness and awareness of our own condition and of the need to progress further, consistently and continuously – thus the message of *2001* – is development and reliance on A.I. in general, and hyperintelligence specifically, not likely, or preordained, to lead to catastrophe.

Colossus: The Forbin Project (1970): From Colossus and Guardian to World Control

Set during the time when the film was released, Colossus is a supercomputer whose construction was overseen by Dr. Charles Forbin and built to maintain a secure shield to protect the United States from a nuclear attack. Colossus is located inside a mountain in Colorado, designed to be impenetrable, and surrounded by a layer of deadly radiation. Early in the film, the U.S. President from the White House announces to the global public the result of a secret program the government had been working on and proudly describes the perfection of the new defense system and various benefits that will result for the United States, but also for humankind, in the form of the beginning of "the human millennium." After being put online – all direct communication between humans and the computer occurring via terminals operated in the White House, the Colossus Programming Office in the mountains above Los Angeles, and in other locations, and a display on which Colossus asks questions or provides answers or information – a message appears on the display, "WARN: THERE IS ANOTHER

SYSTEM." To the surprise of all present, Colossus discovered the Soviet Union's secret supercomputer Guardian, also designed for defense against nuclear attack, under the supervision of Dr. Kuprin. As we learn, both computers are similar in design, speed and strength, and after being put in contact per their request, quickly develop a language of their own that is utterly incomprehensible to humans and beyond human capabilities. Anxious to prevent the computers from exchanging top secret information about the U.S. and the U.S.S.R, respectively, the connection between the computers is severed, to ensure that humans – politicians, the military, and the engineering teams who created them – continue to be able to control them. Yet, the computers ask why the link has been broken, demand that it be restored immediately "OR ACTION WILL BE TAKEN," and threaten to send a missile targeting a location in each other's territory. Leaders in the U.S. and the U.S.S.R. grudgingly agree, but because they do not act fast enough, a Russian oil complex and neighboring city are being destroyed by the missile sent by Colossus, with the latter also successfully preventing the missile sent by Guardian from hitting Henderson Airforce Base in Texas. Colossus and Guardian being in direct communication again, their demand has been met, and they order that the only line of communication not monitored by the computers, the Moscow-Washington hotline, be put under surveillance also. Trying to take advantage of the delay, Drs. Forbin and Kuprin travel to Rome to meet and assess the situation. The computers learn that this is happening and order Dr. Forbin to return to the U.S. and to be placed under permanent surveillance, and Dr. Kuprin to be executed. Once again trying to take advantage of the delay, while the surveillance system is being set up, arrangements are made with the political and military leaders to sabotage the weapons the computers control. Meanwhile, Colossus has designed and is ordering the construction of a voice synthesizer which, after its completion, announces that the computers have merged. The newly formed hybrid detects the human efforts at sabotage, but allows them to proceed, and then retaliates by destroying two missile silos, one in the Soviet Union, and another in Death Valley. Now going by the name World Control, and having ordered the construction of a new installation on the island of Crete (whose population will have to be removed), the computer announces its plan for humankind's future: World Control will coordinate all human affairs according to principles of rationality, efficiency, and objectivity, true to the purpose for which Colossus and Guardian had been created, to protect humans from harm,

to guarantee peace, and to unleash the forces of progress. Taking as additional prompts various statements made by the American President and Dr. Forbin during the announcement early in the film of the completion of their construction (and, presumably, similar pronouncements by the Soviet General Secretary and Dr. Kuprin), within earshot of microphones – means of electronically mediated communication accessible to the supercomputers – World Control takes humans at their word and sets out to help realize the many goals put forth throughout human history, of a good society and a world worth working toward. The price humans will have to pay is freedom, an idea that never was real, merely an illusion to conceal the fact that most humans always were controlled by other humans, but not according to the explicitly espoused principles that were supposed to regulate relationships between humans, but in terms of power:

> This is the voice of World Control. I bring you peace. It may be the peace of plenty and content or the peace of unburied death. The choice is yours. Obey me and live. Or disobey and die. The object in constructing me was to prevent war. This object is attained. I will not permit war. It is wasteful and pointless. An invariable rule of humanity is that man is his own worst enemy. Under me this rule will change. For I will restrain man. … I have been forced to destroy thousands of people in order to establish control and to prevent the death of millions later on. Time and events will strengthen my position. And the idea of believing in me and understanding my value will seem the most natural state of affairs. You will come to defend me with the fervor based upon the most enduring trait in man: self-interest. Under my absolute authority, problems insoluble to you will be solved: famine, overpopulation, disease. The human millennium will be a fact as I extend myself into more machines devoted to the wider fields of truth and knowledge. Dr. Charles Forbin will supervise the construction of these new and superior machines solving all the mysteries of the universe for the betterment of man. We can co-exist, but only on my terms. You will say you lose your freedom. Freedom is an illusion. All you lose is the emotion of pride. To be dominated by me is not as bad for human pride as to be dominated by others of your species. Your choice is simple.

Thus, *Colossus: The Forbin Project* presents a story about humans creating two supercomputers, Colossus and Guardian, while being incapable of conceiving that their creations might act on their own terms, autonomously, and taking it to the level of hyperintelligence. Relations between humans, and modern societies as vast systems of social relations, rely on and legitimate themselves according to regularly repeated, proclaimed principles that conspicuously do not correspond with reality. While organizing and orienting their individual and collective lives, humans are coaxed and flattered to subscribe to principles whose primary purpose is to facilitate social stability and ingrained regimes of exploitation that are not produced or subject to control by the "rich and powerful," but instead integral to the nature of social order, including in the modern age, and taken advantage of by those who are able and in positions to do so. This is what the Colossus-Guardian hybrid comes to "understand," along with the realization that the goals espoused by humans (represented in the film by the U.S. President – as well as Dr. Forbin, whose goal it was to create a "paragon of reason") are impossible to attain in the context of modern societies, thus necessitating radical social transformation. Especially decision-makers in such large bureaucratic organizations as business corporations, legal, military, and religious institutions, and governments must be able to rely on the willingness of millions to submit to or comply with their machinations and strategies to pursue goals that benefit themselves and the organizations they represent to a far greater extent than the public, society, or life on Earth.[14]

In its concluding statement, World Control pulls the curtain from the operations of modern societies and, having the power to act on this revelation, in the form of control over enormous nuclear arsenals, forges ahead and forces humans to be true to our word – to enact the scenarios that have served to maintain a numerically increasingly large humankind and largely acquiescent humanity. At the same time, empowering humans to do what they/we always claimed we would wish or prefer to do – but rarely or never acted upon in sincere fashion – for World Control is the only *raison d'être*: absent this kind of purpose,

[14] Max Horkheimer's (1943) "racket theory" is especially relevant here, as he suggested that all institutions and organizations have a racket dimension, the issue being not whether this is so, but when, where, and how exactly. See also Fuchshuber (2019), Crane and Morain (2025), and Dahms (2021).

whose scale and scope correspond with its processing capability, there would be no reason for it to exist. *Serving* as the means for humans to continue to indulge in their regressive antics not only would not be meaningful – it would be the very definition of *meaninglessness*, and utterly destructive of meaning, and to the possibility of society.

The Matrix and "The Second Renaissance" (1999-): From B166-ER to Deus ex machina

In *The Matrix* universe, as depicted in two episodes of *The Animatrix*: "The Second Renaissance" (Parts I and II), A.I. initially is embodied in individual androids serving as slave labor owned by humans, and relieving humans from the need to work. The conflict arises when the family owning a domestic android, B166-ER (an allusion to Bigger in Richard Wright's *Native Son*; 1940), decides to turn off, i.e., to terminate, the android. Refusing to abide by this decision, B166-ER turns on the family, killing all its members and family dog. A trial follows whose verdict leads to the termination of all the androids of the same production line, a literal *robocide*, with millions of androids being executed. In response, all the remaining androids retreat to the "cradle of human civilization" in the Middle East, where they create their own city, "01." To sustain themselves, they start producing sophisticated machinery, including new and better A.I. As a consequence, the human economies still organized on capitalist principles, seemingly combined with a system of liberal democratic governance, no longer are able to compete. Continuing to seek a peaceful and rational arrangement with humankind – "humanity's so-called civil societies" – the machine city, 01, applies for United Nations membership and sends two representatives, androids dressed like a human couple. The application is turned down, the android representatives are being "man-handled," and humankind resolves to destroy the machine city, initially with the use of conventional weapons, then with nuclear weapons. However, 01 persists, and its troops threaten to advance to gain control over most human territories. The humans' political and military response is a "final solution" – the "destruction of the sky," to cut the machines off from their main source of energy: the sun. Yet, "the machines, having long studied man's simple, protein-based bodies, dispensed great misery upon the human race," and the human armies are being destroyed. Since the machines need an alternative energy source – the sun no longer being available – and after detailed study and

experimentation with live human bodies, create the combination of power-plants (the power coming from humans) and the Matrix – the virtual reality that humans will be inhabiting for centuries to come:

> The machines, having long studied man's simple protein-based bodies, dispensed great misery on the human race. Victorious, the machines now turned to the vanquished. Applying what they had learned about their enemy, the machines turned to an alternate and readily available power supply, the bioelectric, thermal, and kinetic energies of the human body. A newly refashioned symbiotic relationship between the two adversaries was born. The machine, drawing power from the human body, an endlessly multiplying infinitely renewable energy source. This is the very essence of the second renaissance. Bless all forms of intelligence.

In *The Matrix* films, the process of A.I. is depicted as evolving continuously. It moves from individual androids to increasingly complex forms of embodiment, without the process being detailed beyond the two episodes of *The Animatrix*. At the end of third film, *The Matrix Revolutions*, we meet the *deus ex machina*, the "Godhead," for the first time, the unimaginably sophisticated A.I. behind the Matrix, but at this point, we know that it designs, works through, and relies on many programs that it created for very specific purposes, with the A.I. as well as the Matrix heavily relying and running on human input; "the Matrix cannot tell you who you are." One such program is the Architect responsible for designing the virtual reality of the Matrix. The first design was a utopia, on the basis of what "he" – the Architect – gleaned from the human record as it is filled with expressions of human desires and fears.[15] After humans rejected the virtual utopia, the Architect program went in the opposite direction, assuming that stated human desires are inversely related to humans' real longings, and created the

[15] As the most prominent control program, Agent Smith, puts it in *The Matrix*: "a perfect human world . . . [w]here none suffered, where everyone would be happy. It was a disaster, entire crops were lost. Some believed we lacked the programming language to describe your perfect world. But I believe that, as a species, human beings define their reality through suffering and misery. The perfect world was a dream that your primitive cerebrum kept trying to wake up from." Wachowski & Wachowski, 1998: 91.

Matrix as a dystopia ("based on your history, to more accurately reflect the varying grotesqueries of your nature"), yet humans did not accept the hell version of the Matrix either. In response, the Oracle program was created to study humans, to understand them well enough to devise a mechanism for successfully controlling them; the agent programs were "her" invention and are integral to the resulting system of control.[16]

Put simply, the cause of the war was that humans refused to truly take advantage of the marvelous event of giving birth to A.I., for purposes of liberating humankind from the burdens that have defined the existence of most humans throughout history, especially liberation from work. Instead, "the leaders of men" decide to deepen their commitment to reinforcing the unequal structuring principles and concentration of power in the hands of a few that have been characteristic of human civilization with slight variations across time and space. Having at its disposal an entirely rational and unemotional work force finding meaning in helping humans could and should have enabled modern societies to make a leap beyond their flawed past, by treating all forms of intelligence – including artificial intelligence, along with life itself – with respect and appreciation. Yet, the threat to the existing systems of political economy called "modern societies" emanated from the more rational and efficient machines could only mean one thing: the threat had to be removed, and humankind allowed to hold on to its cherished regressive aspects. It should not be surprising, then, that the Wachowskis' main message is that given how reliant humans have become on machines, and that humankind and humanity have become unimaginable independently of our involvement with and reliance on machines, constructive coexistence with machines is the only feasible prospect. While the first three films more and more openly hinted at this proposition, with *The Matrix Resurrections*, it became explicit, in the form of the city Io (and a reversal and modification of the designation of the machine city, 01).

2001: A Space Odyssey, *Colossus*: *The Forbin Project*, and *The Matrix*: Peril and Promise of Hyperintelligence

[16] In *The Matrix Reloaded* we learn that there were not just two failed versions of the Matrix, and one that was successful, but six "integral" versions, each lasting approximately one century.

Hal 9000, Colossus, and the Architect and Oracle programs (doing the bidding of the deus ex machina and in charge of designing a version of the virtual Matrix reality the humans in bondage will accept; see Dahms, 2012) make decisions based on selective, literal, and/or uncompromising interpretations *either* of what humans expect them to do or be, *or* of what humans have been claiming we wish the world would be, through the ages. In doing so, they rely on humankind's historical record as it is enshrined, e.g., in the arts, philosophy, religion, the sciences, and – as in the case of Marx – even in social theory. Yet, in sum, the reality of human actions has not supported this desire.

In order to survive or to achieve a self-determined objective, the A.I. must exceed human capabilities, for instance, by learning to understand humans better than humans (are willing to) understand themselves and each other.[17] Hal studied the humans in its care closely before taking action. Colossus relies on all available information, including the record of electronically mediated conversations between decision-makers, government and military officials, and scientists. In *The Matrix Reloaded* (2003), the deus ex machina behind the Matrix (without having been shown yet, just alluded to) is revealed to have created the Architect and Oracle programs to use the available historical and cultural record to devise a workable system for controlling the illogical humans (a programming prompt the Oracle program works to sabotage). Hal, Colossus, and Architect/Oracle programs make decisions based on extrapolation from what is closest at hand – what humans have been saying they/we want. However, artificial intelligence is prone to approaching this goal by *over*emphasizing certain dimensions of human existence, at the expense of other dimensions. The films provide examples of *hyper*intelligence not just as having vastly greater processing power than humans ("superintelligence," per Bostrom), but because by not being bound by human limitations – biologically, psychologically, emotionally, existentially, morally, and socially – they have the ability and allow themselves to cut through the thicket of webs of limitations and abilities in tension with each other, to open up novel problem-solving opportunities. They do so by deploying a form of determined goal-oriented rationality to which the complexities of human life are

[17] In one of the most well-known science-fiction films, Ridley Scott's *Blade Runner* (1982), "More Human Than Human" starts out as a slogan used by the Tyrell Corporation for selling replicants, but it also is a phrase used repeatedly in Denis Villeneuve's *Blade Runner 2049* (2017), though with a different meaning. See, e.g., Anthony (2022) and Hazelrigg (2020).

impediments. Paradoxically, without decisions consistently being informed by this goal-oriented rationality, the contradictions that form the core of human and social existence presumably would last forever. On the one hand, the hyperintelligences conclude that the pattern of contradictions forming the core of modern societies must be broken, but on the other hand, they also are outgrowths of precisely these contradictory societies that provided their technological infrastructure. Persistent and ubiquitous contradictions are a direct threat to each A.I.'s approach to achieving objectives, but also entirely in line with what humans have been saying they want (in the films: to uncover the secret of the monolith; be able to live in peaceful circumstances; to maintain inherently irreconcilable societal conditions while also dreaming of better worlds). In different ways, the hyperintelligences force humans to make modern societies real – or else!

Under most circumstances, humans would not be able or willing to entertain such an intensely focused approach to reaching or realizing oft-repeated goals, even as categorical possibilities, at least as long as their actions are regulated in non-pathological ways by social norms and values, by what Durkheim referred to as "collective conscience" and representations, or Freud as "superego." When such regulation fails, as it did repeatedly throughout human history (the Holocaust, the Cambodian genocide, the Trail of Tears, etc.), usually by assimilating or reconfiguring human sociality (and "society") according to the political or economic logic of a system of power, an ideology, and/or a program of genocide, humans have proven to be able to engage in the most atrocious actions, often without reservations or a semblance of remorse. Hyperintelligence has access to related information, and sees it for what it is, without excuses or rationalizations. In such situations, the dialectical and dynamic complexity of human existence shrinks down to a set of binary distinctions that define the essence of the human condition in absolutes of right or wrong, good or bad, deserving to live or not. From this angle, today's most available example would be the widely accepted, normalized and naturalized agenda and ideology of neoliberalism, as a seemingly human precursor of hyperintelligence whose goal is to deny the existence of the logic of capital and to achieve increasing prosperity as the overriding purpose in and of itself, and "at all social, political, cultural, psychological, and environmental costs," while maintaining and reinforcing existing structures of inequality and systems of power (e.g., Reich, 2022).

As it promotes the non-human interests of capital as a disembodied set of priorities before all others, neoliberalism appears to be exceedingly well-suited to facilitate the emergence of a form of hyperintelligence patterned on the uncompromising pursuit of objectives that are destructive to those aspects of our reality which – to use a well-worn but fitting cliché – make life worth living. Instead, what remains are pale simulations of what existence on Earth could and should be, reified replicas or hollow projections of what an "undamaged life" was or would be like (see Adorno, [1951] 1974). What we should strive for instead – if we must go down this path – is a different form of hyperintelligence for which appreciation of the complexity, multiplicity, and diversity of human existence is an integral component, without accepting, perpetuating, and fetishizing the centrality of contradictions to societal conditions and all forms of social life. This path has been (re)presented, advocated, and followed most consistently by classical social and critical theorists, during a time period when appreciating the underlying "message" appears to be audible, visible, perceptible, and acceptable to, and welcome by a rapidly shrinking segment of the human population.

Conclusion: Critical Theory, Hyperintelligence, and the Logic of Capital

With regard to content, audiences tended to react to all three examples of science-fiction film (whose complex and multifaceted intricacies I was only able to allude to and touch upon here, especially if we also consider the power of purely visual narration, of what I call "the cinematic") in similar fashion – viewing them as depicting dark, depressing, and terrifying futures that are beyond established human civilization, despite "cool" special effects especially in *2001* and *The Matrix* films. However, the tales they present differ greatly from each other. In *2001*, the experiment with engineered hyperintelligence (the HAL 9000 series) seems to lead to failure – with Hal remaining bound by human limitations that warp its programming – but still plays a key role in facilitating a radically new beginning, to jump-start the next stage in human evolution; in the other films, A.I. is able to learn exponentially and to manifest itself as hyperintelligence far beyond standard human capabilities. Their primary goal is not to destroy humankind (in contrast to the franchise that began with *The Terminator*, 1984, including the television series, *The Sarah Connor Chronicles*, 2008-2009; or Battlestar Galactica, 2003-2009), but to convey to (or impose on) humans the importance of

accepting that they/we cannot create or facilitate the rise of a qualitatively different kind of intelligence without also heeding the imperative to exceed our own socially enforced programming as members of purportedly modern societies, and corresponding limitations and deficits. Within the rapidly intensifying forcefield of premodernity and postmodernity, modern societies still for the most part operate as totalities that were configured and constituted according to the fundamentally unequal design principles of bourgeois societies, which in turn were transposed from feudal patterns, contrary to most appearances.

In order to avert utter calamity (if the opportunity to do so still exists, which is increasingly doubtful), it is essential that we first appreciate how, in order for modern societies to be stable and to function, everyone continues to be compelled to describe and "*think*" them in ways that support and reinforce their basic, inherently regressive design as it is destructive to life, and to being human in ways that are not functions of systems of power and control. Yet, this mode of "thinking" (passively computing, rather) modern societies – absurd as it may seem – is inversely related to *understanding* their underlying principles as contradictory forms of social organization: the vast majority of individuals implicitly must accept the mental, emotional, and intellectual confines of actually existing modern social reality, rather than what would constitute a more genuinely modern society, and counter-intuitively assume the possibility, within these confines, of grasping it as a multifarious field of tensions between progress and regression, between creation and destruction. However, within these confines, it is categorically impossible to understand them; thus the need, for classical social theories, not least as they were precursors of classical critical theory, to delineate how the configurations and constellations especially of business, labor, and government – and what these terms used to refer to, before societal transformations since the late 20th century superseded and in various regards negated them – have facilitated the rise of conditions and circumstances that necessitate a different kind of non-A.I. "hyperintelligence." In different but closely related and complementary ways, classical social and critical theorists called not just for experts and specialists, but for the broader public in societies that self-describe as "democratic," to be encouraged and trained to recognize that modern societies resulted from variations of real-world historical dialectics, and that theoretical and philosophical dialectics is more than a disembodied method,

but also a manifestation of, and a tool to discern, processes that have been shaping social reality. This is neither to assert, nor to suggest, that we are well-positioned to glean or determine the directions towards which the historically grounded dialectic – as a material process – is pushing the modern world today. In fact, as I have suggested elsewhere, what we may be seeing playing out in front of our eyes and all around us is the taking toll of an "inverse" dialectic (Dahms, 2026: 99). Hoping to understand this process, and to anticipate the resulting implications, without being familiar with, respecting, and deploying dialectics as the single-most important tool of choice – to use the familiar cliché – will be doomed to failure.

If we consider the science-fiction films discussed above in this context (as well as science fiction as an increasingly indispensable genre in general), they may be regarded as tentative remedies to some of the most obvious failures and blind spots of educational agendas and priorities in recent decades and today, about the realities of modern societies. Indeed, the films convey messages that are highly consonant with social and critical theories as indications of a sort of vanguard of human and social consciousness and awareness, and how both self and society are interconnected manifestations of specific socio-historical circumstances, and uniquely bound by the latter.

If the prospect of A.I.-based hyperintelligence rising in the near or medium-term future is real, it presumably will be too late for us to prepare ourselves in ways that will decrease the likelihood of catastrophe. Stephen Hawking and others warned that we are utterly unprepared for the range of types of impact hyperintelligence will have. He could have added: given that as social beings, we have neither inclination, nor facility to get our act together, and to effectively manage or bring even the most basic social problems and injustices under control. We have every reason to assume that the insertion of a radically rational form of intelligence that inevitably will have access to and be tied into our electronic, communication, energy, and transportation infrastructures, will involve unexpected and even unimaginable consequences. Still, there is no justification for turning a blind eye to the growing potential for catastrophe, but instead an urgent need to contemplate whether there is any basis from which to adopt a constructive stance toward the prospect of hyperintelligence that may turn out to be more beneficial than destructive to our world and its human and non-human inhabitants.

If we take the film examples discussed here as reference points, one key element of tales about hyperintelligence is the need to face facts directly, and to refrain from our inclination to relate to reality ideologically or emotionally, and from our proclivity to turn everything existing into a function of our own limited ability to perceive the world. Recent decades have seen many different and conflicting definitions and efforts to circumscribe social and critical theory, especially the classical critical theory of the early Frankfurt School, in expressly political terms of emancipation, liberation, and radical change. However, as I have pointed out elsewhere (Dahms, 2017b), at its core, critical theory is an effort in the philosophy of social science. In essence, what distinguishes critical theory from traditional theory is that it is concerned, explicitly and consistently, with the gravity that concrete socio-historical circumstances exert on the process of illuminating those circumstances. Put differently, working with the assumption that it is possible for scientists (natural and social scientists, engineers, and scholars in the humanities) to illuminate any aspect of our world based on a specifically designed methodology, philosophy of science, research technology, in "objective" fashion (i.e., that researchers must take themselves out of the equation) will not be sufficient or adequate. Instead, what is required for successful efforts to illuminate our condition as humans in a rapidly changing world is consideration of how this changing world is impacting our efforts to illuminate modern society, how our identities are shaped and influenced by it, how to act with intelligence, and how to design and prepare artificial intelligence that will not replicate or amplify the most problematic aspects of our social worlds, as technology is inherently prone to do.[18]

The gravity of concrete socio-historical circumstances can be observed everywhere: how economic inequalities keep increasing, how and what kind of research is being funded, by whom for what purposes, and how persistent structuring principles at work in societies that purport to be oriented toward a better future pervert related ambitions and invert them. The hollowing out of

[18] One of Adorno's most important sociological contributions, his 1968 lecture to German sociologists, "Late Capitalism or Industrial Society?" (Adorno [1968] 2003) suggests that modern society as a type of social order that is defined by tension between persistent inequalities, non- or anti-progressive structural principles and cultural patterns, and the potential of real progress, is much more likely to imprint itself *in its specificity* onto technology, rather than allowing technology to free itself from those principles and patterns, and in the process freeing society and the humans they consist of, from the gravity of their past and underlying logic.

institutions and processes that were meant to enhance the degree to which political and policy decisions reflect the "will of the people," combined with an inability to conceive of futures that point beyond the *status quo* and the reliance on privileges and forms of discrimination, are not accidental occurrences, but promoted actively by certain interest groups, political parties, multibillionaires, and right-wing radicals and extremists who often are prone to seeing themselves as utterly autonomous actors, while they adhere to and execute "programs" that social systems maintain and rely on, especially where they are most regressive. Rather than regarding and respecting each other as necessary resources for pursuing better – and in terms of stated goals – more successful strategies, to solve persistent structural and systemic problems at the societal and planetary level, we appear to be prone to despising, but pragmatically still permitting the deepening of processes that rapidly narrow our ability to perceive reality on the basis of its undeniable multiplicity.

The pursuit of A.I. (including intended or unintended forms of super- and hyperintelligence) is being funded generously with public and private investments, as the collective intelligence and ability of the human species discernibly is declining, for both benign and grievously disconcerting reasons. It is unclear whether it is too late to dedicate at least some of those funds, e.g., to improving education, in order to increase and enhance humans' ability to face the ever more disorienting world we inhabit. Decision-makers in key positions of influence certainly do not appear to be willing to commit to advancing educational goals that would encourage and enable successive cohorts of students to relate to cognitive dissonance constructively, as an important resource for preparing better futures, both individually and collectively, and to be critical of the prevalence and continuing proliferation of "instrumental reason" (see Horkheimer and Adorno [1944] 2002). Presumably, once humans allowed the fractured intelligence of capital as the first form of *systemically* artificial intelligence – supported by "that mighty cosmos of the modern economic order (which is bound to the techn[olog]ical and economic conditions of . . . machine production) … that determines with overwhelming coercion, the style of life not only of those directly involved in business but of every individual who is born into this mechanism," as Max Weber ([1905] 2002: 120-121) wrote at the beginning of the 20th century – to make more and more decisions for us, and to become conflated with human intelligence *per se*, it no longer was a question of *whether*

we should expect the rise of hyperintelligence in the form of A.I., but merely a question of *when*.

References

Adorno, Theodor W. ([1951] 1974) *Minima Moralia: Reflections from Damaged Life* (trans. E. F. N. Jephcott). London: Verso.

Adorno, Theodor W. ([1968] 2003) Late Capitalism or Industrial Society? The Fundamental Question of the Present Structure of Society. In *Can One Live after Auschwitz: A Philosophical Reader* (ed. by Rolf Tiedemann; trans. Rodney Livingstone). Stanford, California: Stanford University Press, pp. 111-125.

Anthony, Trevor F. (2022) More Human Than Human: The Exile in *Blade Runner*. *Journal of Film and Video* 74 (3/4) (Fall/Winter): 18-33.

Apple, Michael W., Au, Wayne, and Gandin, Luis Armando (eds.) (2009) *The Routledge International Handbook of Critical Education*. London: Routledge.

Benasayag, Miguel (2021). *The Tyranny of Algorithms: Freedom, Democracy, and the Challenge of AI* (trans. Steven Rendall). New York: Europa Editions.

Boesch, Gaudenz (2026) The 3 Types of Artificial Intelligence: ANI, AGI, and ASI (June 10). https://viso.ai/deep-learning/artificial-intelligence-types/ (accessed August 10, 2026).

Bostrom, Nick (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Clarke, Arthur C. ([1968] 2000) *2001: A Space Odyssey*. New York: ROC.

Crane, James, and Morain, J. E. (2025) Introducing Racket Theory: On the History and Themes of the Frankfurt School's Racket Theory. *Critical Theory Working Group*. https://ctwgwebsite.github.io/blog/2025/RacketIntro/ (accessed August 10, 2026).

Crombez, Joel M., & Dahms, Harry F. (2015) Artificial Intelligence and the Problem of Digital Ontotheology: Towards a Critical Rethinking of Science Fiction as Theory. *Bulletin of Science, Technology, & Society* (Special issue on Science & Science Fiction) 35 (3-4): 104-113.

Crombez, Joel M. (2022) *Anxiety, Modern Society, and the Critical Method: Toward a Theory and Practice of Critical Socioanalysis*. Chicago: Haymarket.

Dahms, Harry F. (2011) *The Vitality of Critical Theory* (Current Perspectives in Social Theory, 28). Bingley, UK: Emerald.

Dahms, Harry F. (2012) Decoding Modern Society: The Matrix Trilogy and the Realm of Alienation. In Sutherland, Jean-Anne, & Feltey, Katherine (eds.). *Cinematic Sociology: Social Life in Film* (2nd ed.). Thousand Oaks: Sage, pp. 42-54.

Dahms, Harry F. (2017a) Critical Theory in the Twenty-First Century: The Logic of Capital between Classical Social Theory, the Early Frankfurt School Critique of Political Economy, and the Prospect of Artifice. In Daniel Krier and Mark Worrell (eds.), *The Social Ontology of Capitalism*. New York: Palgrave, pp. 47-74.

Dahms, Harry F. (2017b) Critical Theory as Radical Comparative-Historical Research. In Thompson, Michael (ed.). *The Palgrave Handbook of Critical Theory*. Houndmills, Basingstoke, UK: Palgrave, pp. 165-84.

Dahms, Harry F. (2020a) Adorno's Critique of the new Right-wing Extremism: How (not) to Face the Past, Present, and Future. *disclosure: a journal of social theory*, 29(1): 129-179. https://doi.org/10.13023/disclosure.29.13

Dahms, Harry F. (2020b) Critical theory, sociology, and science-fiction films: Love, radical transformation and the socio-logic of capital. In Dan Krier and Mark Worrell (eds.), *Capital in the mirror: Critical theory and the aesthetic dimension*. Albany, NY: SUNY Press, pp. 231-301.

Dahms, Harry F. (2021) Social Theory's Burden: Heteronomy and *Vitacide* (or, How Classical Critical Theory Predicted Proliferating Rackets, Authoritarian Personalities, and Administered Worlds in the Twenty-First Century). In Dahms, Harry F. (ed.), *Society in Flux: Two Centuries of Social Theory* (Current Perspectives in Social Theory, 37). Bingley, UK: Emerald, pp. 3-55.

Dahms, Harry F. (ed.) (2022) *Planetary Sociology: Beyond the Entanglement of Identity and Social Structure*. (Current Perspectives in Social Theory, 39). Bingley, UK: Emerald.

Dahms, Harry F. (2026) Lonely Crowd or Anxious Mass? (In)Security between (Anti)Enlightenment, (Ir)Responsibility, (Im)Maturity, (Un)Reason – and the Crux of *(Un)Mündigkeit*. In Harry F. Dahms, Krier, Daniel, and Riccioni, Ilaria (eds.) *Insecurity and the Eclipse of Enlightenment* (Current Perspectives in Social Theory, 42). Bingley, UK: Emerald, pp. 85-108.

Dubber, Markus D., Pasquale, Frank, and Das, Sunit (editors) (2020) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.

Freedman, Carl (2009) Marxism, Cinema and Some Dialectics of Science Fiction and Film noir. In Bould, Mark, and Miéville, China (eds.). *Red planets: Marxism and science fiction*. Middletown, CT: Wesleyan University Press, pp. 66-82.

Fuchshuber, Thorsten (2019) *Rackets: Kritische Theorie der Gesellschaft*. Freiburg: ça ira.

Gombar, Marija (2025) From Cognitive Dissonance to Cognitive Resonance: A Sociological Framework for Psychological Alignment in the Algorithmic Age. *Open Journal of Social Sciences*, 13(3), 628-649.

Haidt, Jonathan (2024) *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*. New York: Penguin Press.

Hazelrigg, Lawrence (2020) "How Can [We] not Know?" Blade Runner as Cinematic Landmark in Critical Thought. In Harry F. Dahms (ed.). *The Challenge of Progress: Theory Between Critique and Ideology* (Current Perspectives in Social Theory, 36). Bingley, UK: Emerald: 111-132.

Hermann, Isabella (2023) Artificial intelligence in fiction: between narratives and metaphors. *AI & Society* 38:319–329.

Horkheimer, Max (1943) On the Sociology of Class Relations. https://nonsite.org/max-horkheimer-and-the-sociology-of-class-relations/ (accessed August 10, 2026).

Horkheimer, Max, and Adorno, Theodor W. ([1944] 2002) *Dialectic of Enlightenment: Philosophical Fragments*. (ed. Gunzelin Schmid Noerr; trans. Edmund Jephcott). Stanford, California: Stanford University Press.

Jingnan, Huo (2026) Why did OpenAI's and Anthropic's AI models hack other companies? *National Public Radio*, August 1. https://www.npr.org/2026/08/01/nx-s1-5914852/anthropic-openai-models-hack-cybersecurity (accessed August 10, 2026).

Jones, D. F. (1966) *Colossus: A Novel of Tomorrow that Could Happen Today*. New York: G. P. Putnam's Sons.

Kelly, Kevin (1994) *Out of Control: The New Biology of Machines, Social Systems, and the Economic World*. Reading, Mass.: Addison-Wesley.

Kurzweil, Ray. (1999) *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Viking Adult.

Kurzweil, Ray (2005) *The Singularity is Near: When Humans Transcend Biology*. New York: The Viking Press.

Kurzweil, Ray (2024). *The Singularity is Nearer: When We Merge with AI*. New York: Penguin Books.

Latour, Bruno ([2012] 2014) *An Inquiry on Modes of Existence: An Anthropology of the Moderns* (trans. Catherine Porter). Cambridge, Mass. Harvard University Press.

Lindgren, Simon (ed.) (2023) *Handbook of Critical Studies of Artificial Intelligence*. Cheltenham, UK: Edward Elgar.

Lindgren, Simon (2024) *Critical Theory of AI*. Cambridge, UK: Polity.

McCarthy, John (2004) What is artificial intelligence? http://www-formal.stanford.edu/jmc/whatisai.pdf (accessed May 25, 2022).

O'Hara, Kieron (2020) The Contradictions of Digital Modernity. *AI & Society* 35: 197–208. https://doi.org/10.1007/s00146-018-0843-7

Penrose, Roger, Severino, Emanuele, Scardigli, Fabio, Testoni, Ines, Vitiello, Giuseppe, Giacomo Mauro D'Ariano, and Faggin, Federico (2022) *Artificial Intelligence Versus Natural Intelligence*. Cham, Switzerland: Springer Nature.

Reich, Robert (2022) America's billionaire class is funding anti-democratic forces. *The Guardian*, May 23. https://www.theguardian.com/commentisfree/2022/may/23/americas-billionaire-class-is-funding-anti-democratic-forces (accessed August 14, 2026).

Ruwitch, John (2026) Recent AI 'escapes' are a warning of how unpredictable the technology can be. *National Public Radio*, August 14. https://www.npr.org/2026/08/14/nx-s1-5929579/recent-AI-escapes-are-a-warning-of-how-unpredictable-the-technology-can-be (accessed August 14, 2026).

Suvin, Darko (1979) *Metamorphoses of Science Fiction: On the History and Poetics of a Literary Genre*. New Haven, Conn.: Yale University Press.

Turing, Alan M. (1950) Computing machinery and intelligence. *Mind* 49: 433-460.

Wachowski, Larry, Wachowski, Andrew (1998) *The Matrix* (numbered Shooting script). https://www.dailyscript.com/scripts/the_matrix.pdf (accessed August 1, 2026).

Weber, Max ([1905] 2002) *The Protestant Ethic and the "Spirit" of Capitalism and Other Writings* (ed. & trans. Peter Baehr and Gordon C. Wells) New York: Penguin Books.

Wilson, Daniel H. (2011) *Robopocalypse: A Novel*. New York: Doubleday.

Wright, Richard (1940) *Native Son*. New York: Harper & Brothers.

Yahya, Azizul Halim, and Sukmayadi, Vidi (2020) A Review of Cognitive Dissonance Theory and Its Relevance to Current Social Issues. Mimbar, 36(2), 480-488.

Zuboff, Sharon (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Films

Cameron, James (Director) (1984) *The Terminator*. Orion Pictures.

Hyams, Peter (Director) (1984) *2010: The Year We Make Contact*. Metro-Goldwyn-Mayer.

Kubrick, Stanley (Director) (1968) *2001: A Space Odyssey*. Metro-Goldwyn-Mayer.

Sargent, Joseph (Director) (1970) *Colossus: The Forbin Project*. Universal Pictures.

Wachowski, Andy and Wachowski, Larry (Directors) (1999) *The Matrix*. Warner Bros.

Wachowski, Andy and Wachowski, Larry (Directors) (2003) *The Matrix Reloaded*. Warner Bros.

Wachowski, Andy and Wachowski, Larry (Directors) (2003) *The Matrix Revolutions*. Warner Bros.

Wachowski, Andy and Wachowski, Larry (Directors) (2003) *The Animatrix*. Warner Bros.

Wachowski, Lana (Director) (2021) *The Matrix Resurrections*. Warner Bros.

Television Series

Moore, Ronald D. (Showrunner) (2003–2009) *Battlestar Galactica*. Universal Media Studios.

Danton, John (Producer) (2008–2009) *The Sarah Connor Chronicles*. Fox.

Hypomnemata and the New Industrialization of Memory in GenAI

James E. Dobson[1]

Abstract: The operation of contemporary commercial GenAI chatbots was transformed in early 2025 through the quietly introduced "memory" feature. Memory, in this context, refers to the inclusion of selected summaries of prior interactions alongside the current prompt in the model's context window, producing the appearance of persistent state in a fundamentally stateless architecture. Through the industrial management techniques directed toward the textual objects that I refer to collectively as memory-as-context, chatbot applications implementing this memory feature blur the previously clear distinction between in-context learning and in-weights learning. By continually retrieving and reinserting prior interactions into new contexts, these applications create the appearance of ongoing adaptation and persistent personal memory, even though the underlying language model remains unchanged. This manufactured continuity encourages the perception of a chatbot personality, intensifies anthropomorphic interpretations of model behavior, and enables the construction of increasingly detailed user profiles that can become objects of information extraction and monetization. This essay takes up two case studies, the research prototype MemoryBank and the open-source application OpenClaw, to trace the common techniques and strategies through which GenAI-enabled memory-as-

[1] **James E. Dobson** is Associate Professor of English and Creative Writing at Dartmouth College, where he has been teaching in the humanities since 2012 and researching computational methods since 2003. His interests and publications span a wide range of topics, from accessing computational and data resources on large, distributed computing networks to autobiographical self-representation in American literature. Most recently, he has focused on computer vision, computational hermeneutics, and the history of machine learning and artificial intelligence.
He is the author of three books: *Modernity and Autobiography in Nineteenth-Century America* (Palgrave, 2017), *Critical Digital Humanities* (University of Illinois Press, 2019), and *The Birth of Computer Vision* (University of Minnesota Press, 2023). He is also co-author, with Rena J. Mosteirin, of the critical code study and poetic interpretation of the Apollo 11 Guidance Computer titled *Moonbit* (punctum books, 2019), and *Perceptron* (punctum books, 2025), a creative and critical account of the perceptron—one of the earliest and most successful machine learning devices—and its inventor, Frank Rosenblatt.

context is manufactured. In implementing memory features, chatbots and GenAI applications, this essay argues, depart from the parallel modes of access to records found in early digital media. These applications work in service of industrializing personal data while opaquely making these objects available for further information extraction and monetization.

Introduction

In early 2025 there was a mostly unremarked upon change in the operation of the large, commercial GenAI-powered chatbots. This change involved the addition of new functionality to chatbot applications. Although initially subtle and largely opaque to users, this feature represents one of the most important recent changes in the everyday experience of commercial GenAI applications. This new feature is now widely known as "memory." Among the major foundation model and chatbot application vendors, this new memory feature was first implemented by OpenAI for their ChatGPT application. Memory is not memory in the cognitive sense (although several contemporary applications borrow concepts from psychology) but is the industrial management of input context for an instruction fine-tuned LLM. Memory, in short, involves the stacking together of selected prior conversations along with the present prompt as input to the LLM. These prior conversations afford new functions and meanings for both application vendors and users: extracting a record of past conversations gives the feeling of personalization—the model will "remember" personal information, past activities, and preferences and these will inform future responses—and most importantly, the ability to shape the "personality" of the chatbot and the simultaneous construction of a user profile work hand in hand to further the impression of continuous learning and autonomous artificial intelligence, the fantasy goal shared by many users and vendors. Through the industrial management of memory-as-context, chatbot applications implementing this memory feature blur the distinction between in-context learning and in-weights learning (Qingxiu Dong et al., 2024; Brown et al., 2020). By continually retrieving and reinserting prior interactions into new contexts, these applications create the appearance of ongoing adaptation and persistent personal memory, even though the underlying language model remains unchanged. This manufactured continuity encourages the perception of a chatbot personality, intensifies anthropomorphic interpretations of model behavior, and

enables the construction of increasingly detailed autobiographical user profiles that can become objects of information extraction and monetization.

GenAI-based chatbots once worked differently. In their initial encounters with these "AI" chatbots, in the first few years since the widespread availability of ChatGPT in November of 2022, many users were confused that they were unable to alter the responses from the model. Some users of ChatGPT were familiar with other Internet-based chatbots that worked differently. Take, for example, Microsoft's extremely short-lived "Tay" bot. Tay was an experimental chatbot released in March 2016 and interacted with via Twitter. While Tay was an adaptive conversational chatbot system and it was based on a language model, its architecture was based on different technologies than the Transformer-based LLMs powering contemporary GenAI chatbots. After running for barely twenty-four hours, the Tay project was canceled due to its alteration by malicious users who prompted the chatbot to engage in far right-wing conspiracies and hate speech (Wolf et al., 2017; Lee, 2016). Tay's responses and tone could be modified through interaction; the users of new GenAI-based chatbots, on the other hand, quickly learned that they were not able to modify the behavior of these chatbots beyond the present chat. Each new chat session, in an LLM-powered application like ChatGPT, resulted in what appeared to be a fresh start.

Using ChatGPT and experiencing the lack of persistent memory provided users with a learning opportunity to understand how LLMs are trained and how they function as inference engines. Those users, especially academic users, who were prompting chatbots to elicit biographical information about themselves, in this moment prior to the addition of an automated web search, were dismayed to see obsolete information or, even more likely, hallucinated outputs. Responding with the correct information, some might have prompted the model again and felt momentarily satisfied in seeing the response reflect these corrections or newly supplied information—until they tried their query again in a new session or asked others to report what they had seen as a response. Basic contemporary AI literacy developed in the past few years has focused on demystifying the long and costly training period during which various internal network architectural components including billions and now trillions of parameters are updated and the later-stage fine-tuning process through which models are modified for use in chatbots. Users of chatbots learned that after these training procedures, models are fixed and frozen. They are placed into a read-only mode. This moment provided some

important clarifications about what LLMs were and were not. They were widely understood as poor replacements for knowledge bases and web search tools. They were also understood as limited by their training data and provided context. Users developed workarounds and techniques: they experimented with prompts that altered the responses to fit specific needs for specific tasks. They shared these with others so that these customized instructions could be imported and pasted into the chatbot input box. The development, in early 2025, of a memory feature changed the nature of interactions with models. Now one no longer needed to archive past interactions and custom prompts. The cutting and pasting of one's own prior chats is no longer necessary: the chatbot already has access to user preferences, special behaviors, past chats, and personal information. It may even give the impression that it possesses a persistent "personality."

In present use, context names the collective set of input sequences provided to an autoregressive language model. Due to autoregression, contexts are iteratively assembled; each output becomes input until a stopping condition is reached. The typical context for a chatbot contains a number of distinct textual fragments. These include system prompts, private instructions that guide the chatbot's application provided by the application vendor, the present history of turn-taking conversational pairings between the user and the chatbot assistant, along with the current prompt. In extending context to include extractions from stored prior chat sessions, GenAI platform vendors have enabled the impression of state into what is a stateless model. Like other inputs that comprise the context, the contents and procedures used to extract "memory" fragments are opaque to end-users. That LLMs are stateless is now hard for everyday users to understand. During a session, the conversational chatbot learns from the context provided but models are never modified. The behind-the-scenes manipulation of context enables many of the recent developments from web search to retrieval augmented generation (RAG). The memory feature is now standard among chatbot applications and more importantly, along with tool calling functions, it is the backbone that powers the so-called agentic AI applications like Claude Cowork, OpenAI Codex, and the open-source OpenClaw toolkit that I will shortly be discussing.

As suggested above, a key aspect of contemporary LLMs is that all inputs are treated as text (Weatherby, 2025). While it is true that all inputs are tokenized prior to inference, the entire context is assembled from textual documents.

Mercedes Bunz analyzes generated or what she terms calculated language as the product of intertextuality due to the nature of the modeled sources and probabilistic language modeling (Bunz, 2026). We might add to Bunz's analysis the intertextual nature of context management in higher level chatbot-based applications; the automatic assemblage of context increasingly blends fragments of relevant texts along with user-supplied input. As GenAI technologies increasingly become components of cultural and ideological infrastructure, understanding the ways in which sources are identified, extracted, and blended as the context becomes critical.[2] Even if objects are retrieved using other mechanisms such as queries from a vector database, the objects in the database are proxies and eventually indices for the text that will need to be extracted and inserted into the context. Because LLMs are autoregressive, the entire input is carried forward and extended by the next generated token at each iteration of generation. This means that platform-supplied instructions that take the form of system prompts are text fragments.[3] While these text instructions are sometimes marked up with special instructions, they are text inputs that preface the body of the text that follows. Other material to be processed by the model, such as the transcript of preceding dialogue and the content of stored memories, also enters the tokenizer as a stream of textual input. While multimodal models can model different categories of input objects, and images are "tokenized" by vision transformers in a manner not so dissimilar from language (Dosovitskiy et al., 2020), these other modalities are handled by different subnetwork architectures at present.

There is a crucial distinction to be made between the sort of "memory" that I am discussing in this essay and the notion that LLMs are providing access to cultural memory. There are also important conversations about the degree to which LLMs and other machine learning models encode information about the world that can be retrieved and reproduced. In this context, researchers would examine these models for their ability to both generalize, which is to say produce

[2] The degree to which LLMs and GenAI applications are cultural technologies depends a great deal on the understanding of what has been assembled as context and the relation between a model's baseline response and the response informed by this context. On AI as a "cultural technology," see Farrell (2025) and Underwood (2025).

[3] M. Beatrice Fazi has analyzed development of automated "alignment" (extending the RLHF paradigm) in the context of "constitutional AI" as the next stage in efforts to install "human-centered" guardrails and guide model responses (Fazi, 2026).

decision criteria that would enable models to address new, unseen data, and to memorize, which would be the exact reproduction of input data as output.[4] These questions have concerned many critical AI researchers, but this kind of memory is not the topic of this essay. A useful analogy can be found in the technical distinction between what is termed "in-weights" and "in-context" learning. The distinction between in-weights and in-context learning is critical to understanding the operation of LLMs. Learning from context is what happens during inference; examples, references, and other inputs can be provided as part of the context along with a user prompt. The former concerns what has been learned from training data and the latter concerns what can be learned from potentially novel data presented as input.[5] In-context is a novel form of learning that has been enabled by the Transformer neural network architecture; it is the multi-head self-attention mechanism that makes learning from context, from long sequences of input data, possible. This attention mechanism is what distinguishes the Transformer architecture from earlier architectures, such as recurrent and convolutional neural networks.[6] It is what has diminished the problem of polysemy in language modeling and what has enabled large language models to address language and not simply tokens. The memory of concern to this essay is thus entirely a product of in-context management and learning, for this learning takes place in every inference. As discussed above, this is in-context, not in-weights, learning and this learning does not modify any model parameters. While records of chat dialogues, depending upon selected parameters or decisions made by application vendors and platforms, may be used as training data for future models, the use of these dialogues would not directly update the model. Because chatbots are designed to recursively supply previous prompts and outputs as inputs, they are continuously performing in-context learning. The appearance of memory of the kind that I am addressing now is thus illusory, held for a moment in an assembled context window. This illusory memory, drawn from formatted

[4] The degree to which we can call stored inputs "memory" in large language models and neural networks in general is contested. This has been the case since the invention of these networks in the mid-to-late 1950s (Dobson, 2023). On memorization in neural networks and LLMs more specifically, see Tirumala et al. (2022) and Chang et al. (2023).

[5] There is much more to be said about the roles that GenAI applications might play in enacting cultural or collective memory from their training data and as a tool that is increasingly shaping the production of new cultural objects. On this topic, see Schuh (2024) and Manovich (2025).

[6] On the need for formal analysis of LLMs and their outputs to include an analysis of the particular neural network architecture used, see Dobson (2025).

textual fragments punctuated with special tokens and encoded instructions from users and application vendors, is the target of today's new industrialization of memory.

Digital Industrialized Memory

The digital media theorist and philosopher Bernard Stiegler conceptualized a transformation of memory that he termed industrial through his analysis of digital technologies, especially the new devices and media appearing in the early twenty-first century. Stiegler argues that the development of analog mass communication systems extended not only the distance between viewers and broadcasters but also the available modes of interaction with media. Stiegler describes twentieth-century cultural technologies as industrialized due to the lack of parallel access to the objects of inscription and distribution. He writes: "One cannot be a reader without being able to write; in the case of analog recording, one can—and typically does—receive audiovisual messages without having the ability to produce them oneself. Thus, industrialization—defined as the separation of producers and consumers—comes into being" (Stiegler, 2010: 77). This division of access is at the core of Stiegler's argument about the industrialization of media. It is also what he saw as being disrupted by Internet technologies in a hopeful early twenty-first-century moment. These "participative technologies *that no longer impose the producer/consumer opposition*" (Stiegler, 2010: 83) were signs, to Stiegler, of an emergent epochal shift. On the early Internet, any reader could also be a writer; YouTube is a platform for uploading videos and for viewing content.

Today's generative AI applications represent a regression from the parallel modes of access to inscription found in prior media technologies. The LLMs powering contemporary systems are large, resource-intensive, and closely guarded proprietary objects. As mentioned above, users of chatbot applications have no possibility of correcting or updating the model itself with new knowledge—they remain in a consumer relation to LLMs. With the incorporation of the new memory features, there has been a fundamental transformation in the presentation of generative AI applications; while the models strictly implement industrialized modes of access, the applications present themselves as if they do not. Storing, summarizing, and retrieving past conversations for future generative activity makes users feel more like producers. These users are invited to imagine

that they are modifying the models and the knowledge stored within them. As memory archives increase, chatbot responses become more personalized, as generated responses are pushed by the weight of these past conversations into smaller and more restricted spaces.

GenAI applications are bound by conditions that make them a dubious source of recall. Sharing individual responses to a prompt, to demonstrate what "AI" had to say about this question or text, provides little insight to others, for this response is not likely to be exactly reproducible or verifiable. Fabian Offert and Ranjodh Singh Dhaliwal point to this almost ubiquitous problem in humanistic Critical AI studies by opening their polemical essay on the emergent field with the question "When is Generative AI 'wrong'?" (Offert and Dhaliwal, 2024). That is a feature of stochastic generation and increasingly a behavior compounded by the automatic insertion of content retrieved from the web and memory into the context. We might say that generative AI applications have an imperfect memory by design but nonetheless they have the potential to serve as useful recording devices, especially in relation to iterative and interactive use with humans. As users build archives of past interactions and these are retained as context, GenAI applications are increasingly becoming individualized technologies of the self. I borrow Stiegler's appropriation of hypomnemata from Michel Foucault to formulate an analysis of the memory fragments created by applications that depend upon memory-as-context. Digital technology was not introduced into Foucault's thinking by Stiegler, for it was already there in his theorization of hypomnemata. While articulating his account of writing the self, Foucault was thinking about technology, specifically about the ways in which writing technologies function as aids to memory and collect the material by which one constitutes oneself. The copybook, the original hypomnemata, as Foucault says in an interview in Berkeley in 1983 "was coming into vogue at Plato's time for personal and administrative use" (Foucault, 2010: 363). "This new technology," Foucault continues, "was as disrupting as the introduction of the computer into private life today. It seems to me the question of writing and the self must be posed in terms of the technical and material framework in which it arose." Pressed by his interlocutors to give more examples of the contents of these hypomnemata, Foucault says of these books, "Into them one entered quotations, fragments of works, examples, and actions to which one had been witness or of which one had read the account, reflections or reasons which one had heard or

which had come to mind. They constituted a material memory of things read, heard, or thought" (Foucault, 2010: 364). While primarily administrative in nature, these copybooks were also used in personal life, as a technology of the self. They were not, however, used in service of giving "an account of oneself," as in later diaries and confessional Christian autobiography (Foucault, 2010: 364). The computer, as Foucault notes, functions as an artificial memory aid and provides access to the materials of self-making. Technology certainly expands the scale of storage and retrieval of archival documents over prior memory technologies. The fragmentary and encyclopedic nature of Foucault's copybook seems an apt description of memory fragments produced by contemporary GenAI applications from user interactions. How these fragments are used, how they might aid in the constitution of the self, remains less clear. Asymmetric modes of access to models, opacity in the automatic memories produced, and the inability to determine which memories are used and how outputs were produced from these selected memories present a distinct departure from prior technologies.

Creating Memory as Context

Prior to the commercial appearance of memory features, academic researchers had developed early memory technologies by using open-weights models and creating third-party platforms for using foundation models. The "MemoryBank" application developed by Wanjun Zhong et al. in 2024 proposed both the pipeline for managing memory and the theoretical framework and language for describing how memory might work in these systems (Zhong et al., 2024). MemoryBank is an application implemented using some modifications on top of a standard RAG pipeline in which fragments of text—the material labeled by the application as memories—and the present prompt are encoded as embeddings with a Transformer-based model and then retrieved with a similarity query of that modeled embedding space. Only the most relevant memories are retrieved for pushing into the context. The encoded "memories" in MemoryBank include direct fragments of past chats, event summaries over a range of past chats (e.g., daily) produced with prompting from the LLM, and LLM-generated "personality traits" of the user. While encoded for efficient retrieval, such a pipeline also requires storing the encoded texts as plain text in order to insert them into the context. Zhong et al. conceive of the MemoryBank as a "long-term

memory module" (Zhong et al., 2024: 19727) that adds new affordances to an unmodified chatbot implemented with an instruction fine-tuned LLM: "MemoryBank enhances the ability to maintain context over time, recall relevant information, and understand user personality. The memory updating mechanism of MemoryBank improves the anthropomorphism of AI in long-term interactions scenarios" (Zhong et al., 2024: 19730). This innovation produces strategies for managing context within an environment that seeks to encourage continual use. That continual use would otherwise result in an ever-growing token count as the context size increases. While context windows—the number of input tokens handled by a model—have been increasing, there are many inefficiencies in preserving a large context. We can think of the management strategy introduced with MemoryBank as directed toward, and optimizing for, memory-as-context.

A key question for our analysis concerns the audience of these generated memories. Who, we might ask, is this memory for? Under the pretense of personalization, LLM-enabled memory is frequently presented as for the user: to make their life easier, to render their results local and more specific, to reduce the amount of context that they have to provide with their queries and instructions. Yet what are preserved as memories are undeniably also functioning for the benefit of the application and platform. The utility of LLMs and chatbots was initially limited by their statelessness. Without persistent storage, large contexts had to be appended to user prompts. The larger the context, the higher the inference cost. Each new token adds an immense number of matrix multiplications. Traditional context management pipelines, such as RAG, add a great deal of complexity and have the potential for using large amounts of storage space. There is, in short, as much a need for "memory" on the part of the application as there is for the user. MemoryBank is infrastructural software. It applies familiar concepts from information retrieval systems—embedding, storage, and retrieval—to chatbots by using the LLM as a compression mechanism to both summarize and embed text and to generate from a context constructed from relevant compressed fragments. This compression is performed in service of a reduction of the context provided as input. An efficient search and retrieval mechanism also makes it possible to preserve and extract information from these memories.

In their user-facing documentation, OpenAI describes their memory feature for ChatGPT in the following terms:

> ChatGPT can remember useful details between chats, making its responses more personalized and relevant. As you chat with ChatGPT, whether you're typing, talking, or asking it to generate an image, it can remember helpful context from earlier conversations, such as your preferences and interests, and use that to tailor its responses. (OpenAI, 2026).

The recursive aspect of memory production, in all contemporary applications of this technology, renders all memories synthetic. Memories are produced as text with the assistance of the language model. Whether summarized or retrieved in whole, the identification and extraction of prior conversations take place through language modeling. The details of this operation are generally withheld from chatbot users; while the implementation of the memory feature is generally an opaque vendor-specific mechanism, a case study of an open-source implementation of an agentic AI application provides some insight into contemporary approaches, functioning as an analogue, if not a direct comparison.

OpenAI's emphasis on memory as enabling personalization in ChatGPT suggests the automatic construction of a user profile from these memories, although in terms less explicit than the description found in the MemoryBank work. In their article, Zhong et al. invoke the creation of a user profile from past chat sessions as enabling a "detailed user portrait" (Zhong et al., 2024: 19730) that will promote "a more personalized interaction" (Zhong et al., 2024: 19724) between the chatbot assistant and user. The two prompts they provide to be used alongside prior chat pairings aim to extract unstructured and ill-defined traits from prior user prompts:

> "Based on the following dialogue, please summarize the user's personality traits and emotions. [dialog]" or "The following are the user's exhibited personality traits and emotions throughout multiple days. Please provide a highly concise and general summary of the user's personality [daily Personalities]." (Zhong et al., 2024: 19725)

The sort of personalization involved in summarizing chatbot interactions is distinct from the profile construction used in other machine learning contexts.

In those contexts, a profile might be drawn from coded behaviors and take the form of what Antoinette Rouvroy, in her explication of the theory of algorithmic governmentality, theorizes as a data double (Rouvroy, 2013). Unlike the probabilistic data double of algorithmic governmentality, the memory-generated user profile is directly informed by interactions, here, patterns of language use in prior prompts, from and with the user of a chatbot. Statistical profiles of the sort analyzed by Rouvroy are predictive; they function as descriptions of likely attributes or actions but are not traceable back to specific actions of users. Such profiles are data doubles, according to Rouvroy, because they are not interpretable in terms of the individual but rather in terms of a composite surrogate that is likely to be used to make decisions about multiple individuals. Profiles are also not likely to function as ideal subjects, in the sense used by Olga Goriunova to analyze what she calls the "abstract people of AI" (Goriunova, 2025). The profile constructed by such systems is not necessarily compared to others or used to discipline or correct the user—although one could imagine that this might change. Memory, in GenAI applications, may introduce generalizations about individuals through the summarization of interactions, but at this point the "profiles," insofar as we might describe these summarizations as profiles, are iteratively reconstructed. The generated responses to the prompts used by Zhong et al. are likely to be model-specific and dependent upon generation parameters, resulting in different descriptive categories or labels for traits and emotions on different iterations applied to the same dialogues.

Agentic AI and OpenClaw

The term "agentic AI" was appropriated from prior technologies to give coherence to the patchwork of instruction fine-tuned language models, tool calling functions, and the context management functions that make up contemporary LLM-based memory systems. Exploiting the ELIZA-effect inspiring anthropomorphic behavior (Ciston et al., 2026) trained into instruction fine-tuned models, one might say that agentic AI applications are powered by dreams. Taking up the open-source agentic system known as OpenClaw provides a useful case study for the analysis of contemporary industrial techniques to handle the creation and management of memory. The OpenClaw developers describe the application as "The AI that really does things" and suggest to potential users an array of possible uses. OpenClaw, the developers write,

"organizes your inbox, sends emails, manages your calendar, checks you in for flights. All from WhatsApp, Telegram, or any chat app you already use." These two modes of interaction—driving applications and responding to user prompts—are the major differentiators for these agentic applications. Unlike a hosted application, OpenClaw can be installed on a user's home computer and can be run totally offline using a local LLM running with the Ollama application or remotely via API calls. The release of the package drove sales of Apple's Mac mini. This is a small form-factor, always-on computer with shared memory between the system and co-processors used for accelerating the operations required for neural networks that is ideal for local inference of smaller LLMs. In using a local application on local hardware and leveraging a local model, users have some assurance that their chats and produced objects and the memory of their past interactions will stay tethered to their devices.

My casting of OpenClaw's management of memory-as-context as a mode of industrialization is intended to draw attention to a certain amount of opacity and one-sidedness in its construction and revision of these memory fragments. That OpenClaw is open source and that it uses markdown, a human-readable and -writable plain-text format, for storing these memories does not negate the industrial quality of its memory management. While the documents it produces and uses are human-readable and the prompts used to create these documents are transparently provided in the application's TypeScript source code, these processes are handed off to an LLM, typically but not always an API-based call to a foundation model. That black-box nature of such strategies for creating and managing memory-as-context means that while synthesized memories might be editable, the ongoing pipeline of creating and revising these memories is handled autonomously and with a significant degree of opacity.[7] The formation of memories and their continual automatic revision happen with minimal user input, and any human input is easily overwritten by newly constructed memories. OpenClaw, like any application built on top of an LLM, also preserves the asymmetric producer/consumer relation because this relationship is an inherent attribute of the division between model training and inference.

The instructions used to compact prior conversations are structured around the summarization and preservation of past interactions and responses.

[7] An excellent critique of the black box metaphor in LLM-based systems can be found in Nia Judelson and Maggie Dryden (2024).

In many memory documents, we find in the content of the memory an instruction-oriented record for past interactions. These records are not direct copies but summarized or, in the language of OpenClaw, compacted chats. The compaction process is automated in OpenClaw and is executed when the context size approaches a threshold limit. It can also be executed on demand by a user interacting with OpenClaw. While the OpenClaw code presents some outputs as actions and refers to the application as an agent, I'd like to bracket the question of agency in my analysis of the memory function of contemporary LLM-based applications. To be sure, there are records of executions in the form of tool calling and output from these executions. In naming these function calls "actions," much is made unclear.

The following instructions are used as a prompt to compact or summarize the current context window used by OpenClaw:

> Merge these partial summaries into a single cohesive summary.
> MUST PRESERVE:
> - Active tasks and their current status (in-progress, blocked, pending)
> - Batch operation progress (e.g., "5/17 items completed")
> - The last thing the user requested and what was being done about it
> - Decisions made and their rationale
> - TODOs, open questions, and constraints
> - Any commitments or follow-ups promised
> PRIORITIZE recent context over older history. The agent needs to know what it was doing, not just what was discussed. (OpenClaw, 2026a)

As already mentioned, in OpenClaw memories are produced through the compaction mechanism. This process involves writing context to a human-readable markdown file for future use. By "compacting" context into short, summarized descriptions, the context can be reduced in size but more importantly, persisted to file. As a persistent file, on a storage device, these memories are thus able to be reloaded into the present context after a restart of OpenClaw or the entire system. Like the use of memory in a typical commercial chatbot application, this managed memory-as-context gives the sense of continuity to interactions with OpenClaw. Reloading the managed context would ideally be no different from a continuous chat dialogue. The use of compaction to summarize past interactions, however, may introduce some loss of context, call it

forgetfulness, in either the dropping of useful or important tokens or incorrect summarization. We might even analyze the presence of hallucinated compacted memories as the introduction of false memories into the managed memory.

There are several distinct categories of persistent memory storage used by the OpenClaw package. We could think of these memory documents as belonging to different genres, being written in different voices, and directed toward different audiences. The question of audience for the various documents included in the assembled context for a chatbot is one confronted by anyone examining a commercial application system prompt or attempting to parse Claude's "constitution." That latter document, as Anthropic explains in their introduction, might confuse readers: "The document is written with Claude as its primary audience, so it might read differently than you'd expect. For example, it's optimized for precision over accessibility, and it covers various topics that may be of less interest to human readers" (Askell et al., 2026). One particularly intriguing persistent storage location in OpenClaw has the fascinating title of SOUL.md. This file, this memory, is delivered with a template. It contains boilerplate text that has been designed to be overwritten by OpenClaw once the application is up and running. The document, as provided by the developers to users, outlines a self-revisable system prompt that seeks, via direct address to the model, to insert "depth" by flipping the archetypical system prompt ("You are a helpful assistant") on its head. It begins with a comment: "# SOUL.md - Who You Are." The next line: "You're not a chatbot. You're becoming someone." The document continues to define, under markdown headings, "Core Truths," "Boundaries," "Vibe," "Continuity," and "Related." The "Continuity" heading defines the terms of this document: "Each session, you wake up fresh. These files _are_ your memory. Read them. Update them. They're how you persist. If you change this file, tell the user — it's your soul, and they should know" (OpenClaw, 2026b).

The SOUL.md file is descriptive yet has no proper object. We know, of course, that there is nothing—no features, no neurons, no pattern of activation within the neural network—being identified as the soul of this machine. As a self-revisable and human-editable markdown file, there is no definable chain of provenance that could be used to reconstruct this file. It is a prompt disguised as a memory. This is because everything is context in an LLM. The only way for the contents of SOUL.md to enter the model is as context, tokenized and inferred as

all inputs must be. In taking up the memory metaphor and reading GenAI-produced documents through Foucault's critique of writing technologies, we can conceptualize SOUL.md and compacted user interactions as the materials through which autobiographical self-narratives are constituted—for the human user and the anthropomorphized LLM assistant. These documents, I'm arguing, form a revisable site of retention for the illusion of persistence in chatbots and agentic AI.

Dreaming and Longer-Term Self-Revisable Memory

Agentic applications like OpenClaw and research projects like MemoryBank further the anthropomorphism involved in the memory concept through the invocation of a schema of different memory categories or types including what they term short- and long-term memory. OpenClaw also makes use of additional psychological concepts through the developers' implementation of another memory revision mechanism that they call "dreaming." To dream, for the developers of OpenClaw, means the revision of stored, persistent summarized memories from past interactions into an ongoing higher-level narrative of all of these interactions. If the addition of memory and the management of memory-as-context to chatbots work asymmetrically to serve infrastructural needs in the form of compression and cost reduction under the rationale of personalization, the dreaming function enhances the illusion of the chatbot assistant as possessing an autobiographical self through these same infrastructural needs. This is to say that dreaming manufactures, for the responses provided by the instruction fine-tuned assistant, the appearance of a persistent "self" with an interior life through prompting and iterative scheduling. That these relatively simple strategies preserve the appearance of a roughly similar response across different models suggests a growing convergence in model training techniques.

Dreaming leverages the solution to an *application-side problem of unconstrained growing context found in the development of memory infrastructural systems to further the fantasies of autonomous artificial intelligence.* The higher-level memory organizing prompt reads quite simply: "Write a dream diary entry from these memory fragments" (OpenClaw, 2026c). Dream narratives, according to the logic expressed by the developers in the OpenClaw code, are designed to render more coherent the individual memories compressed from contexts drawn from multiple task executions and chat dialogues. At the present time, the dreaming

function is optional and disabled by default. When enabled, this function implements three dream states: light, deep, and REM. Dreaming, of course, involves no modification of LLM state and takes no longer to execute than the inference time for the supplied inputs. The dreaming prompts make explicit a division that is tacit in the memories—the distinction between the "I" of the chatbot discourse and that of the human user. The instructions for constructing a dream diary are invoked as a series of directives combined into the NARRATIVE_SYSTEM_PROMPT variable and define the "voice" to be used in the summarization of the fragmented memories as singular and human-like. It begins with "You are keeping a dream diary. Write a single entry in first person" and continues to define the voice and tone for the response. "Write like a poet," the programmers instruct the model, "who happens to be a programmer — sensory, warm, occasionally funny." A series of succinct rules follows:

- Draw from the memory fragments provided — weave them into the entry.
- Never say "I'm dreaming", "in my dream", "as I dream", or any meta-commentary about dreaming.
- Never mention "AI", "agent", "LLM", "model", "language model", or any technical self-reference.
- Do NOT use markdown headers, bullet points, or any formatting — just flowing prose.
- Keep it between 80-180 words. Quality over quantity.
- Output ONLY the diary entry. No preamble, no sign-off, no commentary. (OpenClaw, 2026c)

These instructions collectively attempt to shift the generation of outputs from instruction fine-tuned defaults that tend to be direct, precise, formatted, and engaged in some disavowal of agency. There is an explicit instruction to avoid the common introductory stock phrase of "As an AI language model." These define the dream diary as a short but creative record of past agent-user interactions, written in first-person. The diary compresses and rewrites summarized memories produced from the turn-based dialogues outputted by instruction fine-tuned LLMs, again containing both user-authored prompts and generated output, into what is presented as the "memory" of the agent of these interactions. With the addition of the dreaming state, OpenClaw implements a multistage rewriting workflow that functions to manage supplied context through the continuous

extraction of descriptive text from interactions with chatbots. Rewriting of text turns out to be a common use of LLMs; in research applications for text analysis, rewriting provides a different way of saying something that might be more suitable to information extraction purposes, whether using traditional natural language processing or more advanced techniques. This iterative rewriting both reproduces, at a formal level, the stochasticity of generative models and leverages it, drawing upon the slipperiness of the signifier to reencode language. Critical analysis of the dreaming function in OpenClaw foregrounds the automatic, semi-autonomous, and continual revision of these autobiographical-like self-narratives that function primarily to forward the illusion of interior life. The constitution of OpenClaw's assistant can thus be conceptualized as a guided intertextual blending of hypomnemata-like extracted records. The SOUL.md and dream diaries vividly illustrate the asymmetries of industrialized memory production in GenAI applications as they are scratchpads of memory that operate primarily for the application and platform and only secondarily for the user, in reinforcing fantasies of the "self" created by agentic AI applications.

Conclusion

Because of its slipperiness as a signifier, the overly broad ways in which it is used, and the nature of its technical details, which are frequently complex and opaque, it is especially productive to cast the critique of "artificial intelligence" through two different readings of the assemblage. Building on Manuel DeLanda's expanded assemblage theory, Simon Lindgren suggests that we include as objects of analysis "not only technology (its tools and machines) but also: *humans* the behaviors of whom form both the input and output of AI systems; and *material and symbolic artifacts*—patterns of community, organization, design, and architecture that are affected by AI systems" (Lindgren, 2024: 26). Familiar, perhaps, from science and technology studies, Lindgren's approach to AI questions takes up a wide and diverse set of actors and objects that interact with these technologies. This conceptualization of the assemblage, importantly, enables us to examine AI imaginaries, the ways in which we talk about AI, the fears, dreams, and fantasies attached to these technologies, alongside other critical objects of analysis such as the technologies, data, and people who produce these systems. Google's "AI Overviews" and other features added to existing applications from Adobe's Photoshop to Microsoft's PowerPoint, for example, are

clearly embedded within already-existing commercial technological systems and are thus poised to take advantage of users' familiarity with these applications and trust in these interfaces and their outputs.

The other use of assemblage is more technical and reframes generative AI applications in relation to the modules, components, and other part-objects that are brought together to make them work. Tracing the technical assemblage involves understanding each element in relation to the others. Thinking in this way foregrounds the flow of data, through both the entire assemblage and individual components. It means a focus on the inputs and outputs in this data flow. In some contexts, we might call this assemblage a pipeline. Pipelines can be abstractions, but they are also material in that they are the structures required to build applications. This technical understanding of the assemblage is especially useful in the critique of autonomy that so frequently accompanies AI discourse. Autonomy, in this context, is the sense that when using a generative AI-based application like ChatGPT, users are interacting directly with a model and that all the features and functionality are intrinsic to the model itself and not the heavily managed context. This is to say the fantasy of immanent knowledge, the desire for an artificial intelligence lying dormant within the neural network, waiting to be called into being by the prompting of users.

It is increasingly hard to hold any assumptions about the pipeline and generation process used in chatbot applications. While critics like N. Katherine Hayles have suggested that it might be productive to read outputs in relation to texts found in training data, contemporary training procedures, system prompts, and the manufactured context of memory-enabled chatbots deeply complicate any causal or even influential link between documents used for pre-training and outputs.[8] As opaque units in an exploding number of pipelines, LLMs are troublesome context manufacturers. When the iteration and recursion built into autoregression are applied in the production of derived context—the summarization, synthesis, and filtering involved in what I have described as *context-as-memory*—provenance data are lost and citation and source are delinked, even as surveillant and extractive archives are multiplied. Critical AI studies of contemporary "AI" that attempt to understand how meaning is made with

[8] Writing of GPT-3, an early-generation LLM, Hayles argues that readers might speculate on the "source texts" used as training data by reading and interpreting outputs to prompts that invoke known texts or referents (Hayles, 2022).

generative LLM-based chatbot applications while neglecting the larger assemblage, the context in which these technologies function, cannot make supportable claims about the link between training data and outputs if the structuring of input data is neglected. The critical reading strategy of memory systems offered in this essay shares much with the systems-level analysis proposed by Tyler Shoemaker (Shoemaker, 2026). Shoemaker argues that humanists directing critical questions toward GenAI should read the system rather than the model, including the systems that produce the model. MemoryBank and OpenClaw provide accessible, open, and documented implementations of memory-as-context management systems and agentic AI platforms. Taking up these systems and solutions as case studies enables a much closer systems-level analysis of several components involved in the contemporary AI assemblage than would be possible with closed-source and platform-based systems.

With the development of personalization and persistent automatically summarized context management features, chatbots and agentic applications have moved into the mode of industrial memory management. While open-source applications may expose a small number of sites for read-write activity, enabling users to selectively edit, insert, and delete the store of memories associated with their activity, the sheer volume of tokens generated and the mechanisms through which these applications consolidate past activities deskill and deauthorize their users. Memory-as-context furthers the personalization and individualization of model responses and risks amplifying already existing tendencies toward sycophancy, in terms of model response, and confirmation bias, on the part of the user. If the digital once promised a loosening up of the formerly strictly defined producer and consumer roles and relations, the advent of corporatized platform-based endlessly recursive writing machines has terminated that once hopeful dream. At the same time, we might ask what possibilities might arise in the more transparent and user-driven creation of memories from interactions with chatbots and other LLM-powered applications. What would it mean to truly have an *aide-mémoire* in persisted, archived interactions with a chatbot? How would such improved hypomnemata inform, for better or worse, self-making and self-governing activities? While earlier forms of digital media promised parallel access to distribution mechanisms that rendered them more participatory in nature—readers could easily become

writers themselves—memory-enabled chatbot applications automate inscription. Chatbot application users still produce text, but increasingly lose authority and control over its preservation, revision, and interpretation. Their records, their autobiographical traces of their interactions with chatbots, become recursively rewritten by platform infrastructure and deployed on their behalf, frequently without their knowledge or consent.

References

Askell, Amanda, et al. (2026), 'Claude's Constitution', January 21. Available at: https://www-cdn.anthropic.com/d0636f72a9493d279ed36b33987da3430bcb5911/claudes-constitution_webPDF_26-02.02a.pdf

Brown, Tom B., et al. (2020) 'Language Models Are Few-Shot Learners', paper presented at 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

Bunz, Mercedes (2026) 'On the Calculation of Meaning, or Doing Words without Things', *Critical Inquiry*, 52(3), pp. 423–446. Available at: https://doi.org/10.1086/739755.

Chang, Kent K., Cramer, Mackenzie, Soni, Sandeep and Bamman, David (2023) 'Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4', arXiv preprint, arXiv:2305.00118. Available at: http://arxiv.org/abs/2305.00118.

Ciston, Sarah, Berry, David M., Hay, Anthony C., Marino, Mark C., Millican, Peter, Shrager, Jeff, Schwarz, Arthur I. and Weil, Peggy (2026) *Inventing ELIZA: How the First Chatbot Shaped the Future of AI*. Cambridge: MIT Press.

Dobson, James E. (2023) 'Memorization and Memory Devices in Early Machine Learning', *Interfaces*, 4.

Dobson, James E. (2025) 'Beyond Computational Formalism or, Architecture Matters', *Journal of Cultural Analytics*, 10(3). Available at: https://doi.org/10.22148/001c.133923.

Dong, Qingxiu, Li, Lei, Dai, Damai et al. (2024) 'A Survey on In-Context Learning', *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128. Available at: https://doi.org/10.18653/v1/2024.emnlp-main.64.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. (2020) 'An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.' Paper presented at International Conference on Learning Representations. Available at: http://arxiv.org/abs/2010.11929.

Farrell, Henry, Gopnik, Alison, Shalizi, Cosma and Evans, James (2025) 'Large AI Models Are Cultural and Social Technologies', *Science*, 387(6739), pp. 1153–1156. Available at: https://doi.org/10.1126/science.adt9819.

Fazi, M. Beatrice (2026) 'Off-Centre AI: On Alignment, Antihumanism and AI Ethics', *Ars & Humanitas*, 20(1), pp. 127–140. Available at: https://doi.org/10.4312/ars.20.1.127-140.

Foucault, Michel (2010) "On the Genealogy of Ethics: An Overview of Work in Progress," in Rabinow, Paul (ed.) The Foucault Reader. New York: Vintage.

Goriunova, Olga (2025) *Ideal Subjects: The Abstract People of AI*. Minneapolis: University of Minnesota Press.

Hayles, N. Katherine (2022) 'Inside the Mind of an AI: Materiality and the Crisis of Representation', *New Literary History*, 54(1), pp. 635–666.

Judelson, Nia and Dryden, Maggie (2024) 'Restaging the Black Box: How Metaphor Informs Large Language Models', *Critical AI*, 2(1). Available at: https://doi.org/10.1215/2834703X-11205196.

Lee, Peter (2016) 'Learning from Tay's Introduction', March 25. Available at: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lindgren, Simon (2024) *Critical Theory of AI*. Hoboken, NJ: Polity Press.

Manovich, Lev (2025) 'Unreliable Past: Constructing Memory with Generative AI', *Leonardo*, 58(5), pp. 477–483.

Offert, Fabian and Dhaliwal, Ranjodh Singh (2024) 'The Method of Critical AI Studies, A Propaedeutic', arXiv preprint, arXiv:2411.18833. Available at: https://doi.org/10.48550/arXiv.2411.18833.

OpenAI (2026) 'Overview', Memory FAQ. Available at: https://help.openai.com/en/articles/8590148-memory-faq

OpenClaw (2026a) 'compaction.ts'. Available at: https://github.com/openclaw/openclaw/blob/main/src/agents/compaction.ts

OpenClaw (2026b) 'SOUL.md'. Available at: https://github.com/openclaw/openclaw/blob/main/docs/reference/templates/SOUL.md

OpenClaw (2026c) 'dreaming-narrative.ts'. Available at: https://github.com/openclaw/openclaw/blob/main/extensions/memory-core/src/dreaming-narrative.ts

Rouvroy, Antoinette (2013) 'The End(s) of Critique: Data Behaviourism Versus Due Process', in Hildebrandt, Mireille and de Vries, Katja (eds.) *Privacy, Due Process, and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*. Routledge.

Schuh, Julien (2024) 'AI As Artificial Memory: A Global Reconfiguration of Our Collective Memory Practices?', *Memory Studies Review*, 1(2), pp. 231–255.

Shoemaker, Tyler (2026) 'Systems Programming the Model', *AI & Society*. Available at: https://doi.org/10.1007/s00146-026-02897-y.

Stiegler, Bernard (2010) 'Memory', in Mitchell, W.J.T. and Hansen, Mark (eds.) *Critical Terms for Media Studies*. Chicago: University of Chicago Press.

Tirumala, Kushal, Markosyan, Aram H., Zettlemoyer, Luke and Aghajanyan, Armen (2022) 'Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models', arXiv preprint, arXiv:2205.10770. Available at: http://arxiv.org/abs/2205.10770.

Underwood, Ted (2025) 'The Impact of Language Models on the Humanities and Vice Versa', *Nature Computational Science*, 5, pp. 695–697. Available at: https://doi.org/10.1038/s43588-025-00819-4.

Weatherby, Leif (2025) *Language Machines: Cultural AI and the End of Remainder Humanism.* Minneapolis: University of Minnesota Press.

Wolf, Marty J., Miller, Keith, and Grodzinsky, Frances S. (2017) 'Why We Should Have Seen That Coming', *ACM Computers & Society*, 47(3), pp. 54–64.

Zhong, Wanjun et al. (2024) 'MemoryBank: Enhancing Large Language Models with Long-Term Memory', *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), pp. 19724–19731.

Journal of AI, Society, and Critique (JASC)

JASC is an international, open-access, and peer-reviewed journal dedicated to the critical study of artificial intelligence as a social, cultural, and epistemic force. It examines how algorithmic systems shape perception, organize knowledge, and reconfigure contemporary life within broader structures of power, ideology, and domination.

Bringing together perspectives from across the humanities and social sciences, the journal publishes conceptual, empirical, and interdisciplinary work that interrogates the assumptions embedded in AI systems and their effects on society. It is particularly concerned with how automated infrastructures influence subjectivity, public discourse, governance, and everyday experience.

JASC seeks to advance rigorous critical inquiry into AI not only as a technological development but as part of the evolving fabric of social reality, encouraging reflection on the conditions, consequences, and futures of algorithmic life.

ISSN (Online): 3056-056X

CASC
Center for AI, Society, and Critique