MITIGATING SOCIETAL COGNITIVE OVERLOAD IN THE AGE OF AI: CHALLENGES AND DIRECTIONS

Salem Lahlou
Mohamed bin Zayed University of Artificial Intelligence
salem.lahlou@mbzuai.ac.ae

ABSTRACT

Societal cognitive overload, driven by the deluge of information and complexity in the AI age, poses a critical challenge to human well-being and societal resilience. This paper argues that mitigating cognitive overload is not only essential for improving present-day life but also a crucial prerequisite for navigating the potential risks of advanced AI, including existential threats. We examine how AI exacerbates cognitive overload through various mechanisms, including information proliferation, algorithmic manipulation, automation anxieties, deregulation, and the erosion of meaning. The paper reframes the AI safety debate to center on cognitive overload, highlighting its role as a bridge between near-term harms and long-term risks. It concludes by discussing potential institutional adaptations, research directions, and policy considerations that arise from adopting an overload-resilient perspective on human-AI alignment, suggesting pathways for future exploration rather than prescribing definitive solutions.

1 INTRODUCTION: AI, COGNITIVE OVERLOAD, AND THE LOOMING GOVERNANCE CRISIS

1.1 THE AGE OF SOCIETAL COGNITIVE OVERLOAD: A LOOMING CRISIS

We stand at a precipice. Human societies are increasingly struggling to process the sheer volume and complexity of information in the digital age, a condition dramatically amplified by the rapid proliferation of artificial intelligence (AI). While Toffler (1970) foresaw "future shock" from accelerating change and Eppler & Mengis (2004); Bawden & Robinson (2009) analyzed individual information overload, Byung-Chul Han, in his critique of neoliberalism and technological domination (Han, 2017), argues that contemporary society faces a regime of technological domination that exploits and overwhelms the psyche. This exploitation and overwhelming of the psyche, now dramatically amplified by AI-driven information and complexity, elevates information overload to a systemic crisis: **societal cognitive overload**.

This is a state where individuals, institutions, and even entire governments are overwhelmed, their capacity to make sound decisions eroded by the sheer cognitive demands of navigating AI-driven systems.

This societal cognitive overload manifests across three critical domains:

- **Informational Domain:** The deluge of AI-generated synthetic media (deepfakes (Chesney & Citron, 2019; Tolosana et al., 2020), algorithmically curated filter bubbles (Pariser, 2011) and disinformation campaigns (Vaccari & Chadwick, 2020)) erode shared epistemic ground (Benkler et al., 2018), making it harder to discern truth from falsehood.
- **Moral Domain:** Societies struggle to define fairness in algorithmic systems (Noble, 2018) and grapple with fundamental ethical questions about human agency, responsibility, and values in an age of automation (Carr, 2010; Turkle, 2011).
- **Systemic Domain:** As Tainter (1988) warned, societies risk collapse when complexity outpaces their capacity to manage it. AI-driven interconnectedness (Crawford, 2021) and the potential erosion of human decision-making skills (Kahneman, 2011) exacerbate this systemic fragility.

Critically, this multifaceted cognitive overload not only degrades our capacity to address present-day challenges, across these informational, moral, and systemic domains, but also undermines our ability to grapple with the profound long-term implications of AI. This includes the very real concerns about AI safety and potential existential risks articulated by leading researchers (Bostrom, 2014; Russell, 2019; Bengio et al., 2023; Ord, 2020).

Complementing these concerns about abrupt existential risks, recent research by Kulveit et al. (2025) highlights the equally concerning, yet often overlooked, threat of "gradual disempowerment". They argue that even incremental advancements in AI, by progressively replacing human labor and cognition across crucial societal systems, can lead to a systemic and potentially irreversible erosion of human influence and control, ultimately precipitating a different pathway to existential catastrophe through the slow and subtle undermining of human agency at a societal scale.

1.2 AI AS A DOUBLE-EDGED SWORD

AI technologies act as both cause and potential remedy for this overload:

- **Exacerbation**: Algorithmic manipulation traps users in engagement-driven filter bubbles (Pariser, 2011), automation disrupts labor markets (Acemoglu & Restrepo, 2017; 2020), and opaque systems concentrate power in unaccountable platforms (Srnicek, 2017; Varoufakis, 2023).
- **Mitigation**: AI could enhance human cognition through tools for information filtering (Malone, 2018) or decision support, but only if designed to prioritize societal resilience over profit motives (Norman, 2013).

1.3 BURIED QUESTIONS AND INSTITUTIONAL PARALYSIS

AI forces societies to confront long-deferred questions:

- What level of inequality is permissible in economies where AI concentrates wealth (Acemoglu & Robinson, 2012)?
- How can policymakers govern algorithms when technical complexity overwhelms democratic processes (Green, 2021)?
- What existential risks are acceptable in pursuing artificial general intelligence (Bostrom, 2014; Bengio et al., 2023)?

These dilemmas remain unresolved because cognitive overload paralyzes institutions. As Green (2021) demonstrates, even well-intentioned policies like human oversight of algorithms fail when decision-makers lack the bandwidth to audit complex systems. This creates a *bidirectional misalignment*: overloaded institutions cannot govern AI effectively, while poorly governed AI intensifies societal strain.

1.4 PAPER OUTLINE

This paper argues that mitigating societal cognitive overload is not merely beneficial, but *essential* for responsible AI development and ensuring alignment with human values, particularly when navigating potential existential risks. To support this argument, we will:

- Analyze the key mechanisms by which AI exacerbates societal cognitive overload, spanning informational, economic, and existential dimensions, and highlighting the amplifying roles of deregulation and profit-driven incentives (Section 2).
- Demonstrate how AI, while exacerbating cognitive overload and potentially paralyzing institutions, paradoxically *forces* us to confront critical "buried questions" related to algorithmic fairness, economic inequality, and existential risks that societies can no longer afford to ignore (Section 3).
- Conclude by discussing the potential institutional adaptations, research directions, and policy considerations that emerge from an overload-resilient perspective on human-AI alignment, suggesting pathways for future exploration.

Our central aim is to reframe the AI alignment challenge through the lens of societal cognitive capacity, arguing that reducing overload is a crucial prerequisite for effective governance, ethical AI development, and ultimately, a safer and more human-compatible AI future.

2 AI'S DOUBLE-EDGED SWORD: MECHANISMS OF COGNITIVE OVERLOAD

2.1 EXACERBATION: HOW AI INTENSIFIES SOCIETAL STRAIN

2.1.1 ALGORITHMIC MANIPULATION AND POLARIZATION

The architecture of AI-driven platforms, optimized for engagement, creates self-reinforcing cycles of polarization. Benkler et al. (2018)'s analysis of the 2016 U.S. election demonstrates how networked propaganda exploits cognitive overload: bots and hyper-partisan media outlets flooded social platforms with disinformation, overwhelming users' capacity to discern truth. As Howard et al. (2018) detail, these actors strategically leverage social media algorithms to amplify their messages, using techniques such as bot networks to artificially inflate engagement and targeted advertising to reach specific demographics with tailored disinformation. This "cybernetic loop" of overload and polarization is exacerbated by filter bubbles (Pariser, 2011), where users are trapped in ideological echo chambers. For example, the Brexit referendum saw the deployment of AI-powered microtargeting tools, such as those used by Cambridge Analytica (Cadwalladr, 2019), which leveraged psychographic profiling to deliver tailored misinformation. This exploitation of cognitive biases, including confirmation bias, is further amplified under conditions of cognitive overload. When overloaded, individuals become less capable of engaging the effortful System 2 thinking (Kahneman, 2011) needed for critical evaluation, making them more susceptible to not only initial misinformation, but also to the "continued influence effect" (Lewandowsky et al., 2012), where false information persists even after corrections are presented. This creates a *bidirectional misalignment*: overloaded users are easily swayed by misaligned AI, while these misaligned algorithms further amplify information overload, creating a negative feedback loop. These systems exacerbate societal fragmentation and erode shared reality.

2.1.2 AUTOMATION ANXIETY AND ECONOMIC FRAGILITY

Acemoglu & Restrepo (2020) demonstrates that automation disproportionately displaces low-wage earners in what Autor et al. (2003) characterize as "routine-task intensive" sectors. These sectors, encompassing not only manual labor in manufacturing but also cognitive roles in clerical and administrative work, are particularly vulnerable due to the codifiable and rule-based nature of their tasks, making them readily automatable by AI and robots. The psychological toll is severe: Case & Deaton (2020) links job displacement to rising "deaths of despair" (suicide, substance abuse) in communities hollowed out by automation. Cognitive overload further exacerbates this crisis, as workers facing displacement and the daunting prospect of reskilling must navigate complex and often opaque AI-driven job markets (Webb, 2020). The sheer scale of this reskilling challenge is underscored by OECD studies (Arntz, 2016), which highlight the potential for widespread job displacement across developed economies. Moreover, policymakers themselves, often facing their own cognitive overload amidst rapid technological change and lobbying pressures (Crawford, 2021; Green, 2021), may struggle to enact effective and timely responses. Cognitive overload thus becomes a significant barrier to workers adapting to automation-driven job displacement, hindering their ability to reskill and effectively navigate new, AI-driven job markets. This dynamic contributes to the expansion of what Standing (2011) describes as the "precariat": a growing class facing precarious employment and economic insecurity. The cognitive burden experienced by the precariat, coupled with anxieties about future prospects, not only undermines individual well-being but also diminishes societal resilience and the capacity for a constructive and informed public discourse on AI and automation policy, creating a bidirectional misalignment. This creates a feedback loop where economic precarity erodes societal capacity to govern AI, and unregulated AI deepens inequality.

2.1.3 EROSION OF HUMAN AGENCY

Turkle (2011) documents how increasing reliance on AI-driven social platforms diminishes empathy and self-reflection, leading to a paradoxical sense of being "alone together." This trend towards shallower online engagement resonates with Carr (2010)'s warning about the cognitive consequences

of hyperlinked and algorithmically curated content, which fosters "skimming" over deep reading. Indeed, neuroscientific research (Small et al., 2009) suggests that habitual internet use may alter brain activity patterns, potentially favoring rapid, shallow information processing at the expense of sustained attention and deep analytical skills. The cumulative effect of these trends is a society potentially less equipped for deep deliberation, particularly on complex and long-term issues like AI's existential risks. This erosion of human agency is not a mere side-effect, but arguably baked into the design of many platforms. As Eyal (2014) elucidates in "Hooked", persuasive design principles are strategically employed to maximize user engagement and habit formation, often at the cost of users' conscious control over their attention and cognitive habits. Furthermore, Dennett (2017)'s concept of "competence without comprehension", exemplified by AI systems like based on large language models (Brown et al., 2020; OpenAI, 2023; Guo et al., 2025) generating human-like text without genuine understanding[1], raises deeper concerns about human agency in relation to increasingly sophisticated AI, echoing Vinge (1993)'s warnings about a potential "post-human era" where human control is fundamentally challenged. Cognitive overload, exacerbated by reliance on AI-driven platforms, diminishes human self-reflection and critical thinking. This leads to a **bidirectional misalignment**: humans become less capable of articulating and defending their values in the face of increasingly powerful AI, while AI systems, developed without robust human value input, may drift further from human-compatible goals. This concern about the erosion of human critical thinking skills in the face of increasingly capable AI is echoed in recent commentary, with articles in media outlets like Big Think (Pomeroy, 2025) raising questions about whether over-reliance on AI tools could lead individuals to "take the convenient route of allowing AI to handle our critical thinking", rather than preserving and developing this essential cognitive capacity themselves.

2.1.4 Deregulation, Profit Motives, and Concentration of Power as Cognitive Overload Amplifiers

The exacerbation of societal cognitive overload by AI is not solely a technological phenomenon; it is deeply intertwined with socio-economic and political factors. Current trends in deregulation, the dominance of profit motives, and the increasing concentration of power in the hands of a few large tech corporations (Varoufakis, 2023; Heikkilä, 2023; Verdegem, 2024) act as significant amplifiers of cognitive overload. As governments struggle to keep pace with AI's rapid evolution, the cognitive burden of oversight often leads to de facto or explicit deregulation. This "cognitive offloading" to the market, as discussed by Braithwaite & Drahos (2000) in the context of global business regulation, can be particularly problematic in the AI sector. Fundamentally, the exacerbation of societal cognitive overload by AI is deeply intertwined with prevailing profit motives and the mechanics of the attention economy (Wu, 2016). As Lanier (2018) compellingly argues, the core business model of many dominant tech platforms is not merely about connecting people, but about capturing and relentlessly monetizing user attention. This creates a powerful incentive to design systems that maximize engagement metrics, even if such designs inherently contribute to information overload, algorithmic manipulation, and societal fragmentation, as these often paradoxically increase short-term engagement. This trend is further amplified by the increasing concentration of power and resources within a handful of tech corporations (Zuboff, 2019; Srnicek, 2017). As Khan (2016) powerfully demonstrates in her analysis of Amazon's "antitrust paradox", and as Zuboff (2019) details in her critique of "surveillance capitalism", these entities wield immense power to shape the digital information landscape, control vast data resources, and influence public discourse. This concentrated power directly undermines societal resilience to cognitive overload, as the decisions of these corporations, often driven by opaque algorithms and shareholder value maximization, have far-reaching and largely unaccountable impacts on the information environment and the cognitive well-being of billions. This dynamic reinforces a feedback loop where cognitive overload weakens regulatory capacity, leading to further deregulation and increased corporate power, which in turn intensifies cognitive overload.

2.1.5 Existential Uncertainty and the Cognitive Load of Meaning-Making

Beyond the more readily quantifiable forms of cognitive overload, AI introduces a more subtle yet profound cognitive burden: existential uncertainty and a sense of eroding meaning. As AI systems

[1]The question of whether LLMs genuinely understand language and concepts remains a subject of ongoing philosophical inquiry regarding the nature of understanding and reasoning.

increasingly exhibit capabilities once considered uniquely human—creativity, complex communication, problem-solving—fundamental questions about human identity, purpose, and uniqueness are brought into sharp relief. This growing sense of existential uncertainty is not merely a theoretical concern, but is reflected in public perceptions of AI. A recent National Public Opinion Poll on the Impact of AI (Imagining the Digital Future Center, Elon University, 2024) reveals widespread public anxiety about the societal implications of AI, suggesting a broad societal awareness of the profound and potentially unsettling transformations AI may bring to human life and meaning-making. This can be understood as a form of "existential cognitive load", reflecting the often taxing mental effort and anxiety associated with grappling with these profound questions of meaning and purpose in a world where the boundaries of human and artificial intelligence are becoming increasingly blurred, as explored for example by Yalom (2020)'s work on existential psychotherapy. This challenge to anthropocentric worldviews can create a sense of existential unease and disorientation. The democratization of powerful AI tools, like LLMs, makes these existential questions accessible and relevant to a wider population. While this existential questioning can be a catalyst for philosophical reflection and societal evolution, it also undeniably adds to the overall societal cognitive load, as individuals and communities grapple with redefining their place and purpose in a world increasingly co-inhabited by intelligent non-human agents. This challenge to traditional meaning-making deeply resonates with broader philosophical concerns about the "malaise of modernity", as Taylor et al. (1989) describes, where established sources of identity and value are under strain, leaving individuals and societies searching for new foundations. Furthermore, this existential uncertainty, when combined with the more practical forms of information and task-related cognitive overload, can contribute to a pervasive sense of societal anxiety and a diminished capacity for collective action, potentially exacerbating the decline in social capital and civic engagement highlighted by Putnam (2000).

2.2 Mitigation: Toward Human-Centered AI Design

2.2.1 Context-Aware and Human-Centered Tools for Cognitive Support

AI's potential for cognitive augmentation hinges on designing tools that genuinely enhance human capabilities without exacerbating overload. Malone (2018) envisions AI as "superminds" – collaborative systems that amplify collective intelligence. To realize this, AI tools must be context-aware, adaptive to individual cognitive capacities, and prioritize human agency. Realizing this potential for cognitive augmentation, however, is far from straightforward. The development of truly effective human-centered AI tools for cognitive support faces significant technical and design challenges. It requires not only advanced AI algorithms but also careful consideration of user interface design, human-computer interaction principles, and ethical implications to ensure these tools genuinely empower users without introducing new forms of overload or manipulation.

- Personalized Information Filtering and Sensemaking: AI could offer personalized information filters prioritizing relevance and reducing redundancy, drawing on "universal usability" (Shneiderman, 2000). AI-powered summarization tools based on LLMs could reduce cognitive burden of information processing. However, building effective personalized filters that avoid echo chambers (Gomez-Uribe & Hunt (2015); Nguyen et al. (2014)) remains a technical hurdle. Careful UI/UX design, guided by usability principles (Nielsen, 1994), is crucial.
- Explainable and Transparent AI for Trust and Calibration: Mitigation requires Explainable AI (XAI) providing insights into reasoning (Doshi-Velez & Kim, 2017; Lipton, 2018; Guidotti et al., 2018). XAI helps users understand AI recommendations and calibrate trust. However, creating usable XAI under cognitive overload remains a challenge (Miller (2019); Nielsen (1994)). XAI should offer concise, relevant, actionable insights, aligning with "mixed-initiative" interface design (Horvitz, 1999), where humans and AI collaborate to understand each others' contributions.
- Tools for Collective Deliberation and Consensus Building: Platforms like Polis (Small et al., 2021) show AI's potential for large-scale deliberation. AI tools could identify agreement/disagreement, structure discussions, summarize viewpoints. However, scalability, moderation, UI design for synthesis remain hurdles. Ethical concerns about algorithmic bias and manipulation must be addressed. These tools aim to enhance "collective intelligence" (Woolley et al., 2010).

However, it's crucial to acknowledge the potential pitfalls. AI tools themselves can be designed to be addictive (Alter, 2017), manipulative, or biased. Therefore, human-centered design must be guided by ethical principles and focus on empowering users, not further exploiting their cognitive vulnerabilities. The goal is to create AI that enhances human cognitive capacity and agency, rather than substituting or undermining it. Context-aware AI tools aim to reduce cognitive overload, thereby enhancing human capacity to understand and interact effectively with AI. This contributes to *bidirectional alignment* by empowering humans to better steer AI and ensuring AI systems are designed to be more human-compatible in cognitively demanding environments.

2.2.2 Guardrails Against Cognitive Exploitation and Systemic Overload

Technical solutions alone are insufficient. Mitigating societal cognitive overload also requires robust "guardrails": regulatory frameworks and societal norms that protect against cognitive exploitation and systemic risks. These guardrails must address the bidirectional nature of the problem: preventing AI from exacerbating overload, and ensuring institutions are capable of governing AI effectively. However, establishing effective guardrails against cognitive exploitation and systemic overload is a complex undertaking, fraught with practical, ethical, and political challenges. These guardrails must not only be robust enough to protect against harms but also adaptable to the rapidly evolving AI landscape and sensitive to potential unintended consequences or trade-offs with other societal values, such as innovation and freedom of expression.

- Transparency, Accountability, and Auditing Mandates: Building on the EU AI Act (European Parliament, 2023) and IEEE Ethically Aligned Design (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019), regulations must mandate transparency for "high-risk" AI systems, requiring developers to provide documentation, explainability mechanisms, and undergo independent audits. This reduces the "cognitive auditing burden" on regulators (Green, 2021) by making systems more scrutable. Furthermore, accountability mechanisms are crucial: assigning clear responsibility for harms caused by AI systems, encouraging responsible development and deployment. However, implementing effective transparency, accountability, and auditing mandates for AI systems is far from trivial. As Green (2021) points out, even well-intentioned human oversight policies can falter due to the sheer cognitive burden of auditing complex algorithms. Furthermore, ensuring due process and fairness in algorithmic systems, as explored by Citron & Pasquale (2014), requires careful consideration of procedural mechanisms and regulatory capacity. Building this regulatory capacity, including developing AI auditing tools for regulators themselves and fostering interdisciplinary expertise, is a critical challenge.
- Attention Economy Reforms and "Right to Disconnect" Principles: The attention economy, driven by engagement-maximizing algorithms (Wu, 2016), directly contributes to cognitive overload and societal polarization. Reforms could include limiting the capacity of platforms to algorithmically promote harmful content or filter bubbles, and expanding the right to disconnect (Müller, 2020; Syvertsen, 2020; Baerten et al., 2023) to protect individuals from constant digital connectivity and work encroachment, fostering cognitive restoration. However, regulating the attention economy and implementing the right to disconnect involve navigating complex trade-offs. For example, overly restrictive regulation of algorithmic amplification could raise concerns about freedom of expression and innovation (Napoli, 2019). Similarly, enforcing right to disconnect policies in a globalized and always-on work culture presents practical and cultural challenges. Policy design in this area requires careful balancing of competing values and a nuanced understanding of the digital media ecosystem.
- Labor and Economic Safeguards in the Age of Automation: To mitigate automation anxiety and economic precarity, which exacerbate cognitive overload, policy interventions are needed:
  - Strengthening social safety nets: Universal Basic Income (UBI) or robust unemployment benefits can reduce economic stress and free cognitive resources for adaptation and civic engagement (Danaher, 2019).
  - Investing in reskilling and lifelong learning: Preparing the workforce for the changing job market, but also ensuring reskilling programs are cognitively accessible and effective, rather than adding to overload.

- Promoting human-centered automation: Encouraging AI development that augments human labor rather than solely replacing it, focusing on tasks that are dull, dangerous, or dirty, while preserving meaningful human work.

However, even seemingly beneficial policies like UBI are subject to ongoing debate and scrutiny. As explored in extensive research (Bidadanure, 2019), questions remain about the optimal design, funding mechanisms, and potential societal impacts of UBI, including its effects on work motivation and inflation. Furthermore, ensuring that reskilling programs are truly effective and accessible, and that "human-centered automation" is not merely a rhetorical concept but a practical reality, requires sustained effort and careful policy implementation.

Responsive regulation (Braithwaite & Drahos, 2000), characterized by iterative, evidence-based rules and stakeholder engagement, is crucial for navigating the rapidly evolving AI landscape. "Sandbox" approaches, like Singapore's (OECD, 2024), can allow for controlled experimentation and impact assessment before widespread deployment, fostering a more adaptive and overload-resilient approach to AI governance. Ultimately, these guardrails are essential for ensuring that AI benefits society as a whole, rather than exacerbating existing inequalities and cognitive strains. Regulatory guardrails aim to prevent AI from exacerbating cognitive overload and to ensure institutions have the capacity to govern AI effectively. This is essential for *bidirectional alignment*: reducing societal cognitive overload creates a more stable foundation for developing and deploying AI that is truly aligned with human values, while effective governance mechanisms ensure ongoing alignment as AI evolves.

3 AI as Catalyst: Confronting Buried Societal Questions of Fairness, Inequality, and Risk

While societal cognitive overload presents a significant impediment to addressing complex challenges, the rise of AI paradoxically acts as a catalyst, forcing a long-overdue confrontation with fundamental societal questions that have often remained buried beneath layers of routine and deferred deliberation. In periods of relative societal stability, societies can often postpone grappling with deeply uncomfortable questions about justice, equity, and existential risks. However, the transformative and disruptive power of AI, coupled with the intensifying pressures of cognitive overload, resurfaces these "buried questions" with a new urgency, demanding that we confront them head-on if we are to navigate the AI age responsibly and ethically.

3.1 Defining Justice in Algorithmic Systems: An AI-Driven Imperative

Can justice be mathematically defined for AI? Mittelstadt et al. (2016) argues fairness metrics conflict, revealing fairness is value-laden. Fairness is inherently value-laden, reflecting diverse and often conflicting human priorities. This complexity, often overlooked, is exposed by codifying fairness for AI. Cognitive overload paralyzes ethical discourse, yet amplifies the societal imperative to grapple with these questions. Without consensus on values for **bidirectional alignment**, AI alignment with human interests is impossible, exacerbating societal anxiety and overload. This lack of value consensus undermines AI alignment and erodes public trust in these systems. True justice in the AI age demands participatory design: AI adapting to human values.

3.2 Thresholds of Inequality in Automated Economies

Acemoglu & Robinson (2012)'s influential framework in "Why Nations Fail" distinguishes sharply between "inclusive" and "extractive" institutions. *Extractive institutions*, in their analysis, are designed to concentrate power and wealth in the hands of a narrow elite, hindering broad-based economic progress. In contrast, *inclusive institutions* foster wider participation, protect property rights, and promote competition, creating conditions for innovation and shared prosperity, a distinction that gains critical salience in the context of AI-driven economies. The increasing sophistication and pervasiveness of AI technologies compels us to confront a long-avoided question: What are the acceptable thresholds of economic inequality in societies where AI-driven automation and platform capitalism reshape labor markets and wealth distribution? AI risks entrenching "extractive"

institutions: For instance, AI-driven gig economy platforms increasingly employ algorithmic management to maximize efficiency, often at the cost of worker precarity and reduced benefits (Srnicek, 2017). This dynamic contributes to a "digital precariat" lacking the resources to advocate for systemic change. West (2017)'s research on urban systems, indicating inequality follows power-law distributions, further highlights AI's potential to amplify existing disparities, potentially leading to a future dominated by a "cognitive elite" controlling data and wealth. The looming potential for AI to fundamentally reshape economic structures compels societies to finally grapple with the ethical and political implications of rising inequality and to actively seek pathways toward more inclusive and equitable AI-augmented economies.

3.3 Existential Risk and Societal Deliberation: AI's Unsettling Provocation

Ord (2020) estimates a non-negligible risk of human extinction by 2100, with misaligned AI identified as a leading contributor. While such long-term, low-probability risks can easily be dismissed or deferred in the face of more pressing daily concerns, the sheer scale of potential AI-driven existential threats acts as an unsettling provocation, forcing societies to confront a most fundamental and long-avoided question: What level of existential risk, if any, is ethically and societally acceptable to incur in the pursuit of advanced artificial intelligence? Yet, as we have argued, public deliberation, essential for navigating such complex ethical terrain, is paradoxically paralyzed by "existential fatigue": a cognitive overload subtype where individuals disengage from long-term, seemingly remote risks. The 2017 Asilomar AI Principles (Future of Life Institute (FLI), 2017), while a valuable expert-driven initiative, notably lacked broader public input, reflecting a critical governance gap in addressing AI's most profound long-term implications. To bridge this gap and foster more inclusive deliberation, tools like *Deliberative Polls* that use AI to synthesize citizen inputs and create "cognitive scaffolds" for informed discourse become increasingly vital. The unprecedented nature of AI-driven existential risks, however remote, serves as a profound catalyst, compelling humanity to develop new modes of long-term societal deliberation and governance capable of grappling with threats that transcend immediate human experience and cognitive biases.

3.4 Cognitive Overload as a Catalyst for AI Safety Concerns

The discourse surrounding AI safety, particularly the potential for existential risks, gains critical relevance when viewed through the lens of societal cognitive overload. Leading voices in AI safety, such as Yoshua Bengio (Bengio et al., 2023), Stuart Russell (Russell, 2019), Nick Bostrom (Bostrom, 2014), and Toby Ord (Ord, 2020), have articulated concerns about advanced AI systems becoming misaligned with human values, potentially leading to catastrophic outcomes. However, the capacity of societies to engage thoughtfully with these long-term risks is significantly undermined by cognitive overload. As argued throughout this paper, cognitive overload impairs decision-making at all levels, erodes social cohesion, and fosters instability. These factors not only exacerbate immediate harms from AI but also create a less resilient and less capable global society for navigating the complex and potentially high-stakes challenges posed by advanced AI, including existential risks. Therefore, addressing societal cognitive overload is not merely a matter of improving present-day well-being but is also a crucial prerequisite for effectively mitigating potential long-term AI risks.

4 Conclusion: Reclaiming Cognitive Capacity for Human-AI Alignment

Societal cognitive overload is not merely an individual burden but a systemic crisis that undermines our capacity to govern complex technologies like AI and to ensure their alignment with human values. This paper has argued that mitigating cognitive overload is not just a desirable outcome but a **precondition for achieving meaningful bidirectional human-AI alignment**. While societal cognitive overload presents a formidable challenge, the very act of confronting this crisis, driven by the transformative power of AI, may paradoxically compel societies to finally address long-deferred and fundamental questions about justice, equity, and the future of humanity.

Our analysis of the mechanisms by which AI exacerbates overload points towards several potential pathways for fostering overload-resilient alignment. These may include the development of independent AI ethics and oversight agencies. Such agencies could play a crucial role in conducting

audits of high-risk AI systems to ensure transparency and fairness, developing and promoting ethical guidelines for AI development and deployment, serving as a public resource for information and education on AI risks and benefits, and facilitating public deliberation and stakeholder engagement in AI policy-making. Furthermore, fostering societal cognitive resilience may require investment in centers for digital literacy and cognitive resilience. These centers could be instrumental in developing curricula for schools and lifelong learning programs focused on critical thinking and media literacy, offering training in mindful technology use and overload management, and promoting public awareness campaigns about the attention economy and cognitive well-being. Finally, enhancing participatory AI governance might involve creating and supporting online platforms for informed public deliberation. These platforms could facilitate broader citizen input on AI policy issues, aggregate and synthesize diverse viewpoints to inform policymaking, and provide channels for citizen feedback and oversight of AI systems.

A robust interdisciplinary research agenda becomes essential to deepen our understanding of the cognitive and psychological impacts of AI. Key research areas include quantifying the cognitive and psychological effects of different AI systems, identifying vulnerable populations and specific cognitive vulnerabilities, and developing metrics for measuring societal cognitive overload and resilience. Research should also prioritize developing design principles for human-centered and overload-resilient AI. This includes formulating design guidelines for AI tools that minimize cognitive load and maximize human agency, exploring novel interface designs for human-AI collaboration, and investigating the effectiveness of different XAI techniques in overload contexts. Policy levers for cognitive safeguards could include enhanced regulation of algorithmic transparency and accountability, mandating transparency for high-risk AI systems, requiring audits and impact assessments, and establishing clear lines of accountability. Reforms of the attention economy may be necessary, such as policies aimed at curbing the harms of engagement-maximizing algorithms, regulating algorithmic amplification of harmful content, enacting "right to disconnect" legislation, and promoting alternative media models that prioritize quality information. Finally, robust labor and social safety nets in the age of automation are essential to mitigate automation anxiety and economic precarity, which exacerbate cognitive overload. This may involve strengthening social safety nets, investing in reskilling and lifelong learning, and promoting human-centered automation through appropriate policies.

The future of human-AI alignment hinges not just on technical advancements, but on our ability to cultivate a cognitively sustainable and ethically robust relationship with increasingly powerful artificial intelligence. Only by addressing the challenge of societal cognitive overload can we hope to navigate the complexities of the AI age and safeguard a future where cognitively sustainable human societies flourish alongside increasingly powerful artificial intelligence.

ACKNOWLEDGEMENTS

The author is grateful to Hachem Madmoun, Tom Bosc, Taha Skiredj, Salma André and Harsh Satija for valuable discussions and suggestions.

REFERENCES

Daron Acemoglu and Pascual Restrepo. Robots and jobs: Evidence from us labor markets. *NBER Working Paper Series*, (23285), 2017.

Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2020.

Daron Acemoglu and James A Robinson. *Why nations fail: The origins of power, prosperity, and poverty*. Crown, 2012.

Adam Alter. *Irresistible: The rise of addictive technology and the business of keeping us hooked*. Penguin, 2017.

M Arntz. The risk of automation for jobs in oecd countries: A comparative analysis. 2016.

David H Autor, Frank Levy, and Richard J Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333, 2003.

Robbe Baerten, Valerio De Stefano, and Rik Wouters. The right to disconnect: An ethical perspective. *Available at SSRN 4348227*, 2023.

David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science*, 35(2):180–191, 2009.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, pp. 18, 2023.

Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

Juliana Uhuru Bidadanure. The political theory of universal basic income. *Annual Review of Political Science*, 22(1):481–501, 2019.

Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, 2014.

John Braithwaite and Peter Drahos. *Global business regulation*. Cambridge University Press, 2000.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Carole Cadwalladr. Cambridge analytica a year on:'a lesson in institutional failure'. *The Guardian*, 17, 2019.

Nicholas Carr. *The shallows: What the Internet is doing to our brains*. W. W. Norton & Company, 2010.

Anne Case and Angus Deaton. *Deaths of despair and the future of capitalism*. Princeton University Press, 2020.

Robert Chesney and Danielle Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, 98:147, 2019.

Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

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

John Danaher. *Automation and utopia: Human flourishing in a world without work*. Harvard University Press, 2019.

Daniel C Dennett. *From bacteria to Bach and back: The evolution of minds*. W. W. Norton & Company, 2017.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Martin J Eppler and Jeanne Mengis. The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The information society*, 20(5):325–344, 2004.

European Parliament. EU AI Act: first regulation on artificial intelligence, June 2023. URL `https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence`.

Nir Eyal. *Hooked: How to build habit-forming products*. Penguin, 2014.

Future of Life Institute (FLI). Asilomar ai principles, August 2017. URL `https://futureoflife.org/open-letter/ai-principles/`. Coordinated by FLI and developed at the Beneficial AI 2017 conference.

Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4): 1–19, 2015.

Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law and Security Review*, 2021. doi: 10.1016/j.clsr.2022.105681.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Byung-Chul Han. *Psychopolitics: Neoliberalism and new technologies of power*. Verso Books, 2017.

Melissa Heikkilä. Generative ai risks concentrating big tech's power. here's how to stop it. *MIT Technology Review*, April 2023. URL https://www.technologyreview.com/2023/04/18/1071727/generative-ai-risks-concentrating-big-techs-power-heres-how-to-stop-it/.

Eric Horvitz. Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 159–166, 1999.

Philip N Howard, Samuel Woolley, and Ryan Calo. Algorithms, bots, and political communication in the us 2016 election: The challenge of automated political communication for election law and administration. *Journal of information technology & politics*, 15(2):81–93, 2018.

Imagining the Digital Future Center, Elon University. The national public opinion poll on the impact of ai. Technical report, Imagining the Digital Future Center, Elon University, February 2024. URL https://imaginingthedigitalfuture.org/reports-and-publications/the-impact-of-artificial-intelligence-by-2040/the-national-public-opinion-poll/.

Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.

Lina M Khan. Amazon's antitrust paradox. *Yale lJ*, 126:710, 2016.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint arXiv: 2501.16946*, 2025.

Jaron Lanier. *Ten arguments for deleting your social media accounts right now*. Random House, 2018.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Thomas W Malone. *Superminds: The surprising power of people and computers thinking together*. Little, Brown Spark, 2018.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

Klaus Müller. The right to disconnect. *European Parliamentary Research Service Blog*, 9, 2020.

Philip Napoli. *Social media and the public interest: Media regulation in the disinformation age*. Columbia university press, 2019.

Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pp. 677–686, 2014.

Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

Donald A Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

OECD. Regulatory experimentation: Moving ahead on the agile regulatory governance agenda. Technical report, 2024.

OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Toby Ord. *The precipice: Existential risk and the future of humanity*. Hachette Books, 2020.

Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin Press, 2011.

Ross Pomeroy. Is ai eroding our critical thinking? *Big Think*, January 2025. URL https://bigthink.com/thinking/artificial-intelligence-critical-thinking/.

Robert D Putnam. *Bowling alone: The collapse and revival of American community*. Simon and Schuster, 2000.

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

Ben Shneiderman. Universal usability. *Communications of the ACM*, 43(5):84–91, 2000.

Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi*, 26(2), 2021.

Gary W Small, Teena D Moody, Prabha Siddarth, and Susan Y Bookheimer. Your brain on google: patterns of cerebral activation during internet searching. *The American Journal of Geriatric Psychiatry*, 17(2):116–126, 2009.

Nick Srnicek. *Platform capitalism*. Polity, 2017.

Guy Standing. The precariat: The new dangerous class. *Bloomsbury academic*, 2011.

Trude Syvertsen. Taking back control! the right to disconnect and mobile technology. *New Technology, Work and Employment*, 35(3):249–254, 2020.

Joseph A Tainter. *The collapse of complex societies*. Cambridge university press, 1988.

Charles Taylor et al. *Sources of the self: The making of the modern identity*, volume 1989. Harvard University Press Cambridge, MA, 1989.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. Technical report, IEEE, Piscataway, NJ, 2019. URL https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

Alvin Toffler. Future shock, 1970. *Sydney. Pan*, 1970.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

Sherry Turkle. *Alone together: Why we expect more from technology and less from each other*. Basic Books, 2011.

Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6 (1):2056305120903408, 2020.

Yanis Varoufakis. *Technofeudalism: What killed capitalism*. Random House, 2023.

Pieter Verdegem. Dismantling ai capitalism: the commons as an alternative to the power concentration of big tech. *AI & society*, 39(2):727–737, 2024.

Vernor Vinge. Coming technological singularity: How to survive in the post-human era. *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 1993.

Michael Webb. The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*, 2020.

Geoffrey West. *Scale: The universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies*. Penguin Press, 2017.

Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

Tim Wu. *The attention merchants: The epic scramble to get inside our heads*. Knopf, 2016.

Irvin D Yalom. *Existential psychotherapy*. Hachette UK, 2020.

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