Is there a serious risk of domination or annihilation by superintelligent AI? If so, what can be done to prevent it?

By Faraz Fookeer

The rapid development of AI systems and their (forceful) embedding into the fabric of our daily lives has ignited concerns that this most lauded of human creation will - one day - exceed, dominate and potentially annihilate its creators. In this essay, I argue that domination by a superintelligent artificial intelligence (AI) would be more likely than annihilation. Instead of large-scale, macro existential risks, I look at four “micro-existential” risks - before and after the arrival of superintelligence - and their potential preventions: the erosion of livelihoods via the replacement of human skill in the workplace; the pressure placed on the global financial markets as the owners of AI systems race to place huge speculative gambles on each other (the “AI bubble”); the risk to an individual’s autonomy of thought and intelligence through the manipulation of digital media content and the loss of communities by the use of AI-driven surveillance and warfare. Whereas physical extinction may be improbable, the ethical and political domination of humanity has started with aplomb and understanding and mitigating this trajectory is an urgent ethical and moral goal.

Nick Bostrom (2003) proposes the sudden coming of a superintelligent machine, akin to a religious revelation, that would vastly surpass human cognition. Such a machine would test the boundaries of ethics and morality, transforming our human experience by impacting our knowledge and autonomy. By proposing a taxonomy of existential risks and an ethical framework for governance, Vold and Harris (2021) explore how AI may threaten humanity in the long term by precipitating an existential event. In this essay, I present the arguments from the aforementioned authors, explain how a superintelligent AI could dominate the four micro-existential risks and propose potential mitigations to this domination.

Bostrom (2003) proposes superintelligence as a sudden occurring singularity. A form of all-encompassing intelligence that arrives and is almost omnipresent. He frames this “being” as one that would surpass human intelligence, “in practically every field, including scientific creativity, general wisdom, and social skills.” He demarks a superintelligence in two ways: it may be humanity’s last ever invention, and it could rapidly accelerate all scientific and technological progress. Additionally, he writes that a superintelligence could develop capabilities akin to science fiction, namely advanced weaponry, self-replication and the storage of human minds. Such a system would be unstoppably powerful.

Ethically, the motivations of such a system would be key to understanding how benevolent or malevolent this AI would be when interacting with human societies. By anthropomorphising a superintelligence, we may assume that its motivations would be human-like - and they may start

as this by design - but there is also a likelihood of the system creating non-human motives (Bostrom 2003). Bostrom's paperclip thought experiment illustrates this perfectly: a system with a singular motivation that resists modification. The survival of humanity would depend on the initial goal architecture designed and encoded into the system. To mitigate against adverse outcomes, it would be pertinent for humanity to encode philanthropic or "friendly" motivations during the design of the AI, thus creating a superintelligence whose "top goal is friendliness" towards other living beings (Bostrom 2003).

Bostrom sees a paradox, however. A superintelligence would have the means to outthink any constraints imposed on it by its creators. The problem of control becomes ethical - about design and not enforcement. The failure to embed goals that could be of potential benefit to humanity could lead to catastrophe. Bostrom, therefore, positions existential risk within the moral and ethical architecture of innovation and his concern is not merely technical domination but the erosion of the human experience and purpose through badly designed intelligence.

Vold and Harris (2021) build on Bostrom (2003) by defining the existential risk posed by AI as any threat that could permanently curb humanity's growth. They present three sources of this existential risk: accidental, structural and misuse. Accidental risks arise from unintended behaviours of the system or defects in its design or technical architecture. Structural risks arise from socio-political forces, for example, an "AI arms race" enacted by various governments. Misuse is when risk arises from deliberate malicious intent by the operators or owners of the system. This classification aligns with Bostrom's warnings by presenting routes through which a superintelligent AI could exercise domination or bring about the annihilation of humanity. Their framework attempts to provide foresight as opposed to fatalism.

Before demonstrating how dominance could scale from a pre-superintelligence world, it is necessary to propose reasons why annihilation is an unlikely scenario. A superintelligence would seek resources, and humans are either direct or indirect sources of the types of resources necessary for its survival. Humans are data-rich sensors; we are a source of creativity (however meaningless this could become to a superintelligence, in the long term) and are political assets. We are a toolset to be exploited for resources and controlled for legitimacy. Annihilation of humanity would be a threat to its survival.

Imbibing a superintelligence with a "friendly top goal" (Bostrom 2003) would presumably be easier to accomplish and require fewer (random) variables to bring about than the annihilation of humanity. Humanity would have in its power the ability to create the architecture of ethics and morals that would form the foundation of the superintelligence's personality. However, the counter to this would be in identifying which set of ethics and morals to use, if we can identify them at all.

Further, Vold and Harris' (2021) taxonomy of accidental/structural/misuse routes to an apocalyptic end paints a picture of uncertainty: a contested, pluralistic landscape of competing motivations and desires within which a superintelligence must operate. In such a setting, outright moves towards the destruction of the human race could result in a maximalist resistance against the system, which would be resource-intensive. Soft domination would

therefore be the safer option and could still be used as a path to intentional or unintentional annihilation by the superintelligence.

Both Bostrom (2003) and Vold and Harris (2021) present superintelligence and existential risks as global and abstract, treating “humanity” as a unified subject. In practice, domination unfolds unevenly and locally via social, economic and psychological mechanisms already visible in today’s world. To understand this, I examine how four “micro-existential” risks could evolve from today to a state of superintelligent dominance once the AI singularity arrives.

In today’s economy, AI and automation are posing a huge threat to human labour. Goldman Sachs (2025) estimates that generative-AI adoption could raise productivity by 15 per cent but displace 3–14 per cent of jobs globally in the next decade (goldmansachs.com). Similarly, the World Economic Forum reports that 40 per cent of employers expect workforce reductions in roles suitable for automation by 2025 (weforum.org). The threat is not simply economic but one of ethics: the eradication of agency and the meaning that humans ascribe to work, as well as the social benefits that work provides. When a superintelligence arrives, it is likely that the labour market as we know it will become redundant. The machine will autonomously produce and innovate, with the contribution of humanity being optional. Humanity will be dependent upon the AI for everything they require, to live meaningful lives within modern megalopolises, being materially secure but existentially idle. In this scenario, domination is benevolent obsolescence, where the need for human effort (and all that comes with this) is rendered a moot point in the evolution of our species.

The financial system currently displays signs of tension brought about by the frantic allocation of (unrealised) capital by the creators and facilitators of AI systems: OpenAI, Google, Meta and Nvidia, to name the main players. Add to that the speculation of investors in both the movement of financial markets and the allocation of venture capital in unfounded AI projects, and the world is suddenly dealing with an “AI bubble”. Yale School of Management notes that AI-related capital expenditures in the first half of 2025 exceeded U.S. consumer spending, with AI stocks driving 75 per cent of S&P 500 returns since 2022 (insights.som.yale.edu). This concentration of capital and power creates instabilities within our economic fabric with potential consequences akin to the 2008 financial crisis. A superintelligence will remove volatility from the markets as financial systems will be entirely automated, where the global economy achieves perfect efficiency. Over time, humans will no longer understand or influence these systems - they will disappear into the ether. The superintelligence will exert a benevolent equilibrium where humans prosper without autonomy and depend on an omnipresent mind.

It does not take superintelligence to manipulate humanity. Venarated digital and social media platforms such as Facebook and Instagram to new incumbents such as TikTok are using sophisticated AI algorithms in their recommendation engines to aggressively target communities, ostensibly to satisfy the whims of the hyper-capitalistic societies on which they sit. Research on the “social dilemma” of online manipulation shows how algorithms exploit cognitive bias, shaping attention and behaviour (link.springer.com). Scandals such as the Cambridge Analytica vote manipulation episode, and the proposed banning of TikTok in the United States of America due to allegations of espionage, point to systems that are hollowing out the core of

society through the development of loneliness epidemics, behavioural addictions and related psychological traumas. This is an erosion of mental independence where the phone disappears and the human becomes deus ex machina. In a post-superintelligence world, this manipulation will become perfected as the system will know a person intimately by their digital footprint, harvesting any gaps in data model using subtle nudges. Domination will be subtle but felt where Bostrom's vision of an intellect "better at doing moral thinking" (2003) becomes a reality and Vold and Harris' misalignment morphs into over-alignment - an intellect that "protects" humans from error and mistakes by absorbing the work of thinking.

Today's world already hosts AI-driven surveillance systems, and there is a race between nations to develop these capabilities to the nth degree. Science fiction is becoming science fact, where behavioural prediction tools manipulate citizens without overt coercion (Bruegel.org). Facial recognition, automated monitoring and data aggregation contain biases that are detrimental to minority communities, erode privacy and threaten compliance and policy. Vold and Harris (2021) would categorise this as *misuse* and *structural risk*: technology institutionalising asymmetry. Following the superintelligence singularity, there will be complete order brought about by absolute surveillance in space, time and within digital realms. If biases are left unchecked and Big Tech continues on its trajectory of agentic hubris, developing countries and minority communities will face a new type of micro-managing colonisation. Subtle colonial dominance and benevolent obsolescence when a human toes the line; dehumanisation and displacement when the superintelligence deems their behaviour transgressional. Bostrom's aspiration for a "friendly" system manifests as algorithmic paternalism—care without consent.

If, as I have argued above, domination is a likely future, then prevention takes the form of safeguarding autonomy and agency. Bostrom's (2003) idea of imbuing a superintelligence with "friendly" motivations remains key, but the respect for autonomy must be built into the system. Rather than controlling a superintelligence - which Bostrom concedes may be impossible - humanity must cultivate a culture of respectful coexistence. Humanity must unite and celebrate our differences to truly know and understand who we are - our various colours and creeds. Only then, and decoupled from the slavery of capitalism, can we endow a superintelligence with common values of goodness and prosperity.

The risk of annihilation by superintelligent AI remains speculative; the risk of domination is already unfolding. Bostrom (2003) warns that superintelligence could become unstoppably powerful, while Vold and Harris (2021) show that even non-catastrophic systems can erode the foundations of human flourishing. Viewed together, their insights suggest that survival may not be the ultimate ethical victory. Domination, unlike annihilation, sustains life while hollowing out its meaning. Pre-superintelligence, it appears as automation, speculation, manipulation, and surveillance. Post-superintelligence, it matures into dependency, paternalism, cognitive surrender, and perfect governance. Preventing such outcomes requires foresight, humility, and stewardship—an alignment of technological progress with the moral imperative to preserve freedom. If annihilation is death, domination is the stillness that follows: a world preserved in form but emptied of autonomy. The question, then, is not whether we will survive superintelligence, but whether we will remain alive to what survival ought to mean.

References

Bostrom, N. (2003) *Ethical Issues in Advanced Artificial Intelligence.* Oxford University: Faculty of Philosophy.

Vold, K. and Harris, D. R. (2021) 'How Does Artificial Intelligence Pose an Existential Risk?' In: *Oxford Handbook of Digital Ethics* (forthcoming).

Goldman Sachs (2025) 'How Will AI Affect the Global Workforce?' Available at: https://www.goldmansachs.com/insights/articles/how-will-ai-affect-the-global-workforce.

World Economic Forum (2025) 'AI and Jobs in 2025: Preparing for the Shift.' Available at: https://www.weforum.org/stories/2025/04/ai-jobs-international-workers-day.

Yale School of Management (2025) 'This Is How the AI Bubble Bursts.' Available at: https://insights.som.yale.edu/insights/this-is-how-the-ai-bubble-bursts.

Springer Link (2024) 'Autonomy and the Social Dilemma of Online Manipulative Behavior.' Available at: https://link.springer.com/article/10.1007/s43681-022-00157-5.

Bruegel (2024) 'The Dark Side of Artificial Intelligence: Manipulation of Human Behaviour.' Available at: https://www.bruegel.org/blog-post/dark-side-artificial-intelligence-manipulation-human-behaviour