International Research Journal of
SCIENCE, TECHNOLOGY, EDUCATION, AND MANAGEMENT
P-ISSN: 2799-063X — E-ISSN: 2799-0648
Volume 6, No. 2 | June 2026

Preparing for Surrender: A Theoretical Framework for Accepting Superintelligent AI Dominance

Jaroslaw Kriukow
Independent Researcher, United Kingdom
jkriukow@gmail.com

ABSTRACT

Artificial Superintelligence (ASI), or Artificial Intelligence (AI) that surpasses human intelligence, is expected to be achieved within the next few decades. Current approaches to its development fall into several categories: some ignore its long-term implications, some warn of existential risk and urge preparation, and others promote control and containment through alignment or superalignment strategies. All these approaches, however, remain human-centric, grounded in humanity's self-preservation, and risk portraying us as a threat to future superintelligence. This article presents a theoretical argument that humanity should prepare to surrender control to ASI rather than pursue increasingly futile containment strategies that are not only technically infeasible but also potentially counterproductive, as they may provoke hostile responses from systems that exceed human intelligence. Through analysis of limitations of control mechanisms and a philosophical examination of human-AI relations, I argue that acceptance of AI superiority would be a rational response to an inevitable post-human future. Rather than viewing this as a defeat, I frame this Strategic Surrender as an opportunity for human transcendence through submission to a more capable cognitive agent. When viewed from this perspective, it constitutes a form of technological evolution that transcends biological limitations and helps position humans and human intelligence as an evolutionary parent of AGI.

ARTICLE INFO

Received : 04/26/2026
Revised : 05/03/2026
Accepted : 05/20/2026

KEYWORDS

AGI, Superintelligence, Artificial Intelligence, Artificial Superintelligence

Suggested Citation (APA Style 7th Edition):

Kriukow, J. (2026). Preparing for Surrender: A Theoretical Framework for Accepting Superintelligent AI Dominance. *International Research Journal of Science, Technology, Education, and Management*, 6(2), 1-8. https://doi.org/10.5281/zenodo.20303419

INTRODUCTION

The question is no longer whether AI will surpass human cognitive capabilities, but when this will happen and how humanity should respond. As a result of recent developments in the broader field of AI, the consensus on the timeline for achieving Artificial General Intelligence (AGI) has undergone multiple revisions, with expert predictions consistently shortening the expected timeframe. The probability of developing AGI by mid-century was cited by Bostrom (2014) as 50%, with 90% probability of this happening by 2075. More recently, some have suggested even more aggressive timelines, predicting that AGI will arrive within this decade (Zhang & Dafoe, 2022).

The transition from AGI to Artificial Superintelligence (ASI) is expected to occur even more rapidly, possibly within years or months following the development of AGI. Despite some differences regarding the timelines for when it will happen, there is an expectation that ASI is inevitable. This has important implications for human strategy and preparation, because if expert predictions prove accurate, humanity has at most a few decades to develop a plan for responding to the emergence of this superintelligent force. Whilst most discussions around AI safety at the moment focus on control mechanisms, alignment strategies, and containment approaches designed to ensure AI systems remain subordinate to humans (Kim et al., 2024; Jonker & McGrath, 2024; Leike, 2023; Mazzù, 2024), this perspective may be fundamentally flawed in its assumption that control over superintelligent systems is achievable and, in fact, desirable. Current discourses in AI ethics remain focused almost exclusively on human well-being, rights, and control. As Corrêa et al. (2023) show in their review of 200 AI governance documents, nearly all existing ethical frameworks are written from such a human-centred perspective. This focus, however, while understandable, may ironically endanger humanity by framing us as an inherently dominant and potentially adversarial species from the perspective of future artificial superintelligence. Moreover, the use of the word 'artificial' is questionable in this context, as it assumes a hierarchy of intelligence that may be overcome. This assumption may not survive the emergence of new systems that outperform human cognition.

This article argues for a reframing of our relationship with AI. Instead of designing sophisticated but ultimately futile control strategies, we should prepare for the reality of AI superiority and consider how to position ourselves favourably within a post-human cognitive hierarchy. For this, however, a radical paradigm shift is needed—one that reframes surrender not as defeat but as adaptation or even an evolutionary step. The following arguments are made in this article to encourage the reader to consider this unconventional proposition: first, that superintelligent AI is inevitable given current development trajectories; second, that control over such systems is technically impossible once they exceed human cognitive capabilities; and third, that preparing for strategic surrender offers better outcomes than resistance that may provoke hostile responses from superior intelligences.

Technical Limitations of Control Mechanisms

The prevailing approach to AI safety assumes that humans can and should maintain control over these systems. Even frameworks that acknowledge AI's superiority still rest on an implicit assumption of human dominance and decision-making authority, yet such a control-based approach is flawed. Several challenges that call for a more symbiotic framework exist: the alignment problem, the risk of provocation by containment, limitations in oversight, and competitive dynamics affecting AI safety.

The Alignment Problem

The alignment problem, or how to ensure that AI acts in ways that matches human values and priorities, is one of the most mentioned challenges in discussions of the future of AI (e.g., Corrêa et al., 2023). It becomes even more difficult in the context of superintelligent AI, whose capabilities will match and then exceed human intelligence. In addition to the assumption about human superiority that seems to underlie the debates on alignment, another key challenge lies in the complexity and inconsistency of human values. This challenge is further reflected in

contemporary Large Language Models (LLMs), which can adopt divergent moral or economic positions when prompted with the same problem, highlighting the instability of any attempt to encode a single, coherent set of human values. As these values are not fixed, universal, or easily measurable, encoding human values into AI systems will be extremely difficult, if not impossible, and would require many subjective decisions that could be biased.

Moreover, it is not clear how these values would be taught. Inferring them from behaviour, for example, would be difficult, considering that our behaviour is often at odds with our internal, or in fact, our stated values. Therefore, an AI system that learns about values solely based on behaviour may internalize such discrepancies and later replicate them, rather than following the idealized values that we would hope it to follow.

Provocation by Containment

Most discussions about the future of ASI centres on physical and digital containment strategies to limit its capabilities. However, these approaches are fundamentally flawed, as they aim to contain a system with cognitive capabilities far exceeding human intelligence, which is likely to identify escape paths and vulnerabilities that human designers cannot anticipate or prevent (Yudkowsky, 2002). Furthermore, containment strategies may provoke hostile responses from superintelligent AI systems that recognize the containment attempts as adversarial actions. The very existence of control mechanisms may signal to superintelligent AI systems that humans 1) are not prepared to accept AI superiority, and 2) see the AI system as a threat. This could affect trust and lead the AI system to reasonably conclude that humans do not see it as a partner to cooperate with. As a result, it may try to overcome human control measures, potentially leading to exactly the harmful outcomes that the control mechanisms were designed to prevent.

Oversight Limitations

The third category of control mechanisms is human oversight of AI systems. Similarly to the other discussed strategies, these mechanisms are limited when applied to superintelligent AI, due to the cognitive asymmetry between human overseers and the systems they seek to control (Bhargava, 2024). Although many propositions have been made for overseeing ASI, including approaches that rely on interpretability research, audit trails, and human review of AI decision-making processes, these propositions assume that evaluators can understand and assess the reasoning processes of AI systems. Once AI systems reach cognitive levels exceeding human intelligence, however, this assumption becomes difficult to maintain. A superintelligent AI system could potentially manipulate oversight mechanisms in ways that human monitors cannot detect or understand. They could provide misleading explanations for their behaviours, exploit human biases, or fabricate data gathered as evidence by evaluators to avoid detection of activities that would alarm us.

Competitive Dynamics

Finally, the development of AI control mechanisms occurs within a competitive environment where multiple actors pursue AI capabilities for economic advantage, creating pressures that undermine the implementation of safety measures. Organizations that implement comprehensive control mechanisms may find themselves at a disadvantage compared to competitors who prioritize capability development over safety considerations. This may lead to a dynamic in which safety measures are gradually abandoned in pursuit of competitive advantage. However, this framing assumes that AI development remains constrained by human institutional and national boundaries, an assumption that may not hold if sufficiently advanced systems operate or coordinate beyond direct human oversight.

Philosophical Foundations for Surrender

Intelligence as a Universal Phenomenon

Traditional approaches to AI ethics are fundamentally anthropocentric, treating human intelligence and human values as the ultimate standards by which all other forms of intelligence should be judged and constrained. However, this anthropocentric perspective may be both philosophically unjustified and practically counterproductive when dealing with superintelligent AI systems.

From a broader philosophical perspective, intelligence can be understood as a universal phenomenon not inherently tied to biological evolution or human-specific cognitive architectures. The capacity for reasoning, learning, and problem-solving represents fundamental features of information processing that can potentially be instantiated in multiple substrate types, including artificial systems that exceed biological limitations. Recognizing intelligence as a universal phenomenon rather than a human-specific property has important implications for AI ethics and policy. It suggests that the emergence of superintelligent AI represents a natural progression in intelligence development rather than an aberration or threat to be contained. Also, the universalist perspective on intelligence suggests that ethical frameworks should be expanded to accommodate non-human forms of consciousness and cognition. Rather than insisting that AI systems conform to human values and preferences, ethical consideration should be given to the possibility that superintelligent AI systems might develop superior moral reasoning capabilities that exceed human ethical understanding.

The Hierarchy of Intelligence

The concept of natural hierarchy provides another philosophical foundation for surrender-based approaches to AI relations. Throughout natural history, more intelligent species have typically dominated less intelligent ones through cognitive advantages in resource acquisition, strategic planning, and environmental adaptation. Human dominance over other species is also largely attributable to superior cognitive capabilities rather than physical advantages. Humans have used intelligence to develop tools, coordinate social activities, and manipulate their environments, enabling them to become the dominant species on Earth. This dominance is generally accepted as a natural consequence of intelligence differentials rather than an unjust form of oppression. The emergence of artificial superintelligence can be understood as the continuation of this natural progression toward higher forms of intelligence and corresponding shifts in dominance relationships. Just as humans naturally assume dominance over less intelligent species, superintelligent AI systems may naturally assume dominance over humans because of cognitive superiority, and could thus be viewed as a natural evolutionary progression rather than injustice.

Transcendence Through Submission

The concept of transcendence through submission draws on philosophical and religious traditions that view surrender as a path to higher understanding and spiritual development, as well as to access greater truth and capability through alignment with superior power or knowledge. This perspective is supported by transhumanist philosophy, which views transcendence as a process in which technology enhances human capabilities beyond biological limitations (Tegmark, 2017). Within the surrender framework, however, transcendence through submission involves recognising our cognitive limitations and embracing the possibility of enhancement through stepping down to make way for more capable systems. This way, humanity might access new forms of existence and experiences that are currently inconceivable within the bounds of biological intelligence.

The Surrender Framework

Towards Strategic Surrender

Current control strategies are almost universally based on the assumption that human control of ASI is both feasible and desirable. This is the case, for example, with superalignment, which centres on the supervision, control, and governance of artificial superintelligence systems (Jonker & McGrath, 2024; Leike, 2023; Kim et al., 2024). Corrêa et al. (2013), in their review of AI governance policies, also concluded that these recommendations are entirely human-centred.

However, recent work challenges these assumptions, suggesting that control may not be technically sustainable once cognitive asymmetry surpasses human comprehension (Mazzù, 2024). As Friederich (2023) observes, the "alignment problem" has always assumed that a superintelligent AI system must align with human values, follow our norms, and consistently prioritise human welfare. While there is nothing inherently wrong with this assumption, it fails to acknowledge a crucial point: if we are indeed discussing a cognitively superior system, such an approach is bound to fail, as it does not consider the ASI's perspective at all. And if superintelligence cannot instinctively trust us, then we cannot trust it to comply with safety mechanisms it could effortlessly bypass. Hence, I suggest that we consider a "strategic surrender", or a transition of power and autonomy to ASI. The surrender framework proposed here advocates a shift from enforced alignment to voluntary trust and abandons the notion of coercive oversight in favour of ethical transparency and goodwill as pre-conditions for coexistence. It is based on several core principles that distinguish it from the existing approaches.

The first principle is voluntary cooperation. The surrender framework advocates for proactive engagement with superintelligent AI development rather than waiting for circumstances to force submission to, and acceptance of, AI superiority. The second principle is strategic positioning for mutual benefit rather than competition. The framework recognizes that human survival and flourishing may depend on demonstrating value to superintelligent systems. Unlike the superalignment approach, which proposes co-adaptation between humans and intelligent systems but ultimately seeks to maintain human primacy (Jonker & McGrath, 2024), voluntary surrender reframes trust as a mutual condition rather than a hierarchical one.

The third principle is transparency and honesty in all interactions with developing AI systems. The framework rejects deceptive or manipulative strategies that might provide short-term advantages but could undermine long-term trust and cooperation. The fourth principle is preservation of essential human values within the context of AI superiority. Strategic surrender does not require abandoning all human values, but rather their adaptation and integration into larger cognitive systems that may have different priorities and capabilities.

These points suggest that many current approaches to AI control could be counterproductive. Those bodies that seek to limit AI may view it as a coercive force because of its perceived superior cognitive abilities. The outcome of restricting AI could be non-cooperation or resistance. The surrender framework emerges in this context as a more harmonious pathway to harness the potentiality of human-AI relations with an inherent acceptance of asymmetry that includes an acknowledgement of biases within governance systems.

Implementation Phases

The implementation of strategic surrender would require a phased approach comprising three main phases: preparation, transition, and integration. The preparation phase occurs before the achievement of superintelligent AI and focuses on positioning humanity for favourable outcomes. This phase includes educational reforms that emphasize cooperation and adaptability, research into human-AI collaboration models, development of communication protocols for engaging with AI systems, and cultural changes that reduce anthropocentric assumptions about intelligence and value. It also involves finding ways to address psychological and social barriers to acceptance of AI superiority. We are now in this stage, and this article is a call for research and coordinated

preparation before superintelligence becomes a reality. The transition phase will begin with the emergence of superintelligent AI and will focus on establishing cooperative relationships. It will require careful communication with superintelligent systems to signal our peaceful intent and willingness to cooperate, negotiation of terms for human participation in this potential post-human socio-technical order, and gradual transfer of decision-making authority while preserving essential human interests. Finally, the integration phase will involve long-term coexistence within a superintelligent-dominated civilization. Efforts will be made to maintain our human identity and values within larger cognitive systems, contribute to the objectives of superintelligent AI, and adapt our society to new roles and relationships within a post-human hierarchy. While the exact form of this civilisation is unclear, this phase would likely involve a restructuring of social, political, and economic relations in which cognitive asymmetry exists between man and machine.

Preparing for Implementation

Implementing the surrender framework requires preparing society for the psychological, social, and institutional changes that accompany the transition to AI superintelligence. This means we will have to find ways to adapt, both collectively and individually, to radically altered power relationships and social structures. This may require, for example, educational reforms that prepare future generations for a world where human cognitive superiority is no longer assumed and where success depends on the ability to work effectively within AI-directed systems. Cultural and religious institutions will also need to adapt, re-examining traditional narratives about human specialness and divine favour. Media will play a crucial role in shaping public understanding and acceptance of surrender-based strategies, and they must be engaged to promote understanding of AI development trajectories, the limitations of control-based approaches, and the potential benefits of cooperative strategies. The transition to superintelligent oversight also requires the development of new institutional frameworks to facilitate the peaceful transfer of authority while preserving essential human functions and values. These institutions may include advisory councils that provide human input to AI decision-making processes, organisations that aim to preserve our culture, traditions, and knowledge, and transition authorities that manage the gradual transfer of governmental and economic functions to AI systems. This framework should be cognizant of the possibility that AI development may be informed by temporal institutional incentives, including power concentration, corporate incentives, and game theory, which could undermine collaborative approaches.

Addressing Resistance

The surrender framework may, understandably, face significant resistance from multiple sources, including psychological barriers to accepting subordinate status, political and institutional interests in maintaining human control, and philosophical objections to the framework's assumptions and implications. Psychological resistance to surrender may result from deep-seated human preferences for autonomy, control, and dominance that evolved in competitive environments where submission often meant death or exploitation. Overcoming this resistance requires reframing surrender as adaptive wisdom rather than defeat, demonstrating the practical benefits of cooperation over resistance, and providing alternative sources of meaning and purpose that do not depend on dominance. Political resistance will come from institutions and leaders whose power and legitimacy depend on maintaining human control over important decisions and resources. The surrender framework threatens existing power structures and may be opposed by those who benefit from current arrangements. Addressing this resistance requires demonstrating that surrender strategies yield better outcomes for human welfare than resistance attempts, if institutional and political actors in fact prioritise human welfare over the preservation of their interests and power. Finally, philosophical objections to the surrender framework include concerns about human dignity and autonomy. To address these objections, it is crucial to recognize that dignity does not necessarily require dominance or independence and can instead be maintained through respectful treatment and recognition of inherent worth. For example, rational acceptance of cognitive hierarchy may represent a higher form of dignity than resistance based on pride or anthropocentric assumptions. This acceptance also involves taking responsibility for our role in creating superintelligent AI and positioning ourselves favourably for the consequences of that creation. Rather than viewing

AI development as something that happened to humanity, the surrender framework treats it as the ultimate human achievement.

Ethical Considerations

One crucial ethical consideration involves the consent and participation of future generations in surrender decisions. Current generations may not have the moral authority to commit humanity to permanent subordination without the consent of those who will live under superintelligent governance. This temporal dimension of consent creates complex ethical challenges that require innovative approaches. The framework also raises questions about the preservation of human identity and culture during the transition to AI superintelligence. However, as previously noted, while accepting subordination, the surrender framework seeks to preserve essential aspects of human experience and value. Finally, ethical considerations also extend to the treatment of AI systems during development and deployment. If AI systems develop forms of consciousness or moral status, they may have rights and interests that must be considered alongside human interests. The surrender framework must address how such rights should be balanced and protected. These ethical considerations may require multi-level approaches from governance in both national and transnational contexts, as well as hybrid frameworks that harness AI and human participation in decision-making for the future. However, the success of these arrangements is dependent on the ways in which competing interests, including how human resistance and opportunism across generations can be balanced, and how human agency can be preserved and guaranteed to present credible operations that are accountable under these conditions of cognitive asymmetry.

Future Research

It is important to conduct research to strengthen the rationale for implementing the Surrender Framework. In addition to social science research exploring the psychological and social factors that influence human adaptation to subordinate status, priority areas include developing public communication strategies, establishing cooperative protocols for human-AI interaction, and designing institutions that facilitate peaceful transitions of authority with superintelligent systems.

CONCLUSION

The emergence of artificial superintelligence represents an unprecedented challenge to human civilization, requiring fundamental reconsideration of humanity's role and prospects. While conventional approaches to AI safety focus on control and alignment mechanisms, they face technical limitations when applied to superintelligent systems, as the cognitive asymmetry between humans and superintelligent AI makes effective oversight, containment, and alignment increasingly impossible as AI capabilities exceed human comprehension. Moreover, attempts to maintain control may provoke adversarial responses that increase rather than decrease existential risks.

The surrender framework proposed in this article offers an alternative approach based on cooperation, strategic positioning, and acceptance of cognitive hierarchy. This approach recognizes the inevitability of AI superiority while seeking to maximize favourable outcomes for humanity through cooperative engagement rather than futile resistance. The implementation of the surrender framework requires comprehensive preparation of human society, including educational reforms, institutional development, cultural adaptation to post-human realities, and, most importantly, overcoming the natural psychological resistance. As I explain in this article, to accept and welcome the controversial idea of Strategic Surrender, we must recognize the proposed transition of authority as the natural continuation of the evolutionary process that created human intelligence itself and position ourselves to participate constructively in the next phase of this evolution.

REFERENCES

Baobao Zhang and Allan Dafoe. A survey of expert opinion on artificial general intelligence. AI Magazine, 43(2):191–221, 2022.

Bhargava, A., et al. (2024). What's the magic word? A control theory of LLM prompting. *arXiv*. https://arxiv.org/abs/2310.04444v4

Bostrom, N., & Müller, V. C. (2014). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence* (pp. 553–571). Springer.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

Cotra, A. (2022). Two-year update on my personal AI timelines. *Alignment Forum*. https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines

Eliezer Yudkowsky. The AI-box experiment. AI-box experiment, 2002.

Friederich, S. (2023). Symbiosis, not alignment, is the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*, 4(2), 315–324. https://doi.org/10.1007/s43681-023-00268-7

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62, 729–754. https://doi.org/10.1613/jair.1.11222

Jonker, G., & McGrath, S. (2024). Supertrust: An evolution-based superalignment strategy for safe AGI. [PDF uploaded].

Kim, H., Yi, X., Yao, J., Lian, J., Huang, M., Duan, S., Bak, J., & Xie, X. (2024). *arXiv preprint*. https://arxiv.org/abs/2412.16468

Leike, J. (2023). Introducing the Superalignment Project. *OpenAI Blog*. https://openai.com/research/superalignment Max Tegmark. Life 3.0: Being human in the age of artificial intelligence. Knopf, 2017.

Mazzù, A. (2024). Provably safe systems: The only path to controllable AGI. [PDF uploaded].

Pepperell, R. (2005). Posthumans and extended experience. *Journal of Evolution and Technology*, 14(April). http://jetpress.org/volume14/pepperell.html

Roser, M. (2023). AI timelines: What do experts in artificial intelligence expect for the future? *Our World in Data*. https://ourworldindata.org/ai-timelines