A Soft Landing into the Singularity: Mediated Control through AGI-Produced Algorithmic Solutions

Luca Rivelli*

2025

Abstract

This paper examines the tension between the growing algorithmic control in safety-critical societal contexts—motivated by human cognitive fallibility—and the rise of probabilistic types of AI, primarily in the form of Large Language Models (LLMs). Although both human cognition and LLMs exhibit inherent uncertainty and occasional unreliability, some futurist visions of the "Singularity" paradoxically advocate relinquishing control of the main societal processes–including critical ones–to these probabilistic AI agents, heightening the risks of a resulting unpredictable or "whimsical" governance. As an alternative, a "mediated control" framework is proposed here: a more prudent alternative wherein LLM-AGIs are strategically employed as "meta-programmers" to design sophisticated–but fundamentally deterministic–algorithms and procedures, or, in general, powerful rule-based solutions. It is these algorithms or procedures, executed on classical computing infrastructure and under human oversight, the systems to be deployed–based on human deliberative decision processes–as the actual controllers of critical systems and processes. This constitutes a way to harness AGI creativity for algorithmic innovation while maintaining essential reliability, predictability, and human accountability of the processes controlled by the algorithms so produced. The framework emphasizes a division of labor between the LLM-AGI and the algorithms it devises, a rigorous verification and validation protocols as conditions for safe algorithm generation, and a mediated application of the algorithms. Such an approach is not a guaranteed solution to the challenges of advanced AI, but–it is argued–it offers a more human-aligned, risk-mitigated, and ultimately more beneficial path towards integrating AGI into societal governance, possibly leading to a safer future, while preserving essential domains of human freedom and agency.

*FISPPA Department, University of Padua, Padova, Italy. (luca.rivelli@gmail.com).

1 Introduction: the tendency to algorithmic regimentation and the AGI horizon

The contemporary world is increasingly characterized by what we could call the "algorithmicization" of human behavior[1], a trend most visibly manifested in safety-critical domains: from surgical checklists to protocols in air traffic control, healthcare, nuclear power plants, space activities[2], we observe a pervasive drive to the regimentation of human action through adoption of predefined Standard Operating Procedures (SOP)[3], which are step-by-step rule-based standardized procedures. This reflects a recognition of inherent human fallibility in complex, high-stakes scenarios: human performance is susceptible to errors from fatigue, stress, cognitive biases, and limitations in working memory and attention[4]. Standardization and regimentation through algorithms–broadly, step-by-step instructions–are seen as crucial to enhance system reliability and safety of critical processes, minimizing variability and ensuring consistent adherence to best practices either through direct execution by digital computers, or, in other cases, by human operators who are compelled to strictly conform to the logic of the rule-based procedure.

However, as we rely more and more on algorithms to manage existing complexities, humanity stands on the cusp of Artificial General Intelligence (AGI). Recent advancements in AI, particularly Large Language Models (LLMs), fuel both excitement and apprehension about machines possibly matching or surpassing human cognitive capabilities. The *technological singularity* envisions AIs capable of recursively self-improving, leading to an intelligence explosion and societal transformations of unprecedented scale, with possible profound benefits but also significant existential risks[5].

This creates a fundamental tension: we increasingly use algorithms or rule-based procedures to control safety-critical processes because human behavior is inherently *probabilistic*: adaptable but error-prone. Yet, LLMs, constituting the current leading path to AGI, are *also* fundamentally probabilistic, based on statistical inference: while implemented through complex algorithms at a lower level, the behavior of LLMs emerges from statistical inference and prediction, resulting in outputs that are often creative and insightful, but also demonstrably prone to inconsistencies, "hallucinations", and a lack of guaranteed reliability across diverse and novel contexts[6]. This raises a critical question: *If both human cognition and the leading architectures for advanced AI are fundamentally*

[1] Gawande (2009).

[2] Terblanche, Fowler & Sibbald (2008), Barbir & Bezzola (2012), Clay-Williams & Colligan (2015), Guy, Kerstein & Brennan (2022), Hales, Borchard, Schwappach, Butterworth (2010), U.S. Nuclear Regulatory Commission (1981), GUIDE, DRAFT SAFETY (2020), Degani & Wiener (1991), Dismukes & Berman (2010).

[3] Akyar (2012).

[4] Reason (1990), Staal (2004), Almarzouki (2024).

[5] Vinge (1993), Kurzweil (2005), Bostrom (2014).

[6] As section 2 will elaborate.

probabilistic, is it safe, or even prudent, to directly entrust the governance of increasingly complex and critical societal systems to these inherently probabilistic artificial general intelligences?. This paper argues that the sole prudent answer to this question is, evidently, *no*.

The alternative solution we put forth is: instead of directly ceding control to LLM-based AGIs, we propose a *mediated control* approach. Our central thesis is that we should leverage the remarkable capabilities of LLM-AGIs not as direct controllers themselves, but as powerful *meta-programmers*, specialized tools capable of devising and coding highly sophisticated yet fundamentally *deterministic algorithms*. It is these AGI-designed algorithms, rather than the probabilistic LLMs themselves, that should be entrusted with the governance of critical societal and socioeconomic processes. This path to the singularity, characterized by a "mediated control" given to the AGI, offers–we contend–a possible "soft landing" into a future shaped by advanced AI: a future that harnesses the transformative potential of AGI while simultaneously safeguarding human agency, ensuring system reliability, and mitigating the inherent risks of relinquishing control to inherently probabilistic intelligences.

2 The probabilistic turn: human and machine cognition as inferential and uncertain

To better ground the argument for mediated algorithmic control, we first show that both human and advanced artificial intelligence, specifically Large Language Models (LLMs), operate as fundamentally *probabilistic cognitive systems*. This section will delineate the evidence for this "probabilistic turn" in our understanding of both human and machine intelligence, highlighting the inherent uncertainty and inferential nature of their cognitive processes.

2.1 Human cognition as inherently probabilistic

Historically, cognitive science often assumed the mind was a *deterministic rule-based system*. However, contemporary cognitive science and neuroscience converge on a different picture: the human brain is seen as an *inference machine* operating under pervasive uncertainty[7]. The *Bayesian Brain Hypothesis* and its related framework of *predictive processing* posit that the brain engages in *probabilistic inference*, constantly updating its internal models of the world based on noisy and ambiguous sensory inputs[8].

According to this view, perception is not passive reception, but active *hypothesis testing*. The brain generates probabilistic predictions about the causes of sensory inputs, comparing the predictions to actual sensory information and continuously updating models to minimize *prediction error*[9]. This inferential process is

[7] Gregory (1980).
[8] Friston (2010), Clark (2013), Hohwy (2013).
[9] Friston (2010).

inherently probabilistic due to an uncertain world, noisy inputs, and limited computational resources. Thus, human cognition is not about achieving perfect certainty, but about making the best possible inferences given limited and imperfect information.

Behavioral economics and studies of *bounded rationality* and *cognitive biases* provide further evidence[10] of the probabilistic nature of human cognition. Limited cognitive resources (*bounded rationality*) lead to heuristics and simplified models of the world to make decisions efficiently. While adaptive, these methods highlight *cognitive biases*–predictable deviations from normative rationality–showing the inherent approximations and probabilistic nature of human decision-making.

From the standpoint of the neurosciences, brain architecture performs a kind of probabilistic processing. Neural activity exhibits *neuronal variability* and *noise*[11]. Neurons fire probabilistically, and neural populations encode information through *probabilistic population codes*[12], where information is represented in the statistical distribution of neural firing rates rather than in precise, deterministic signals. This inherent neural variability suggests that probabilistic computation is not merely a high-level description of cognition, but is deeply embedded in the fundamental mechanisms of brain function.

2.2 LLMs as probabilistic systems

Contemporary *Large Language Models (LLMs)*, despite being implemented as algorithms, function fundamentally as probabilistic systems in their observable behavior and output generation. While–since they are computer programs–LLMs are deterministic at the lowest level of abstraction, at a higher level they are describable as operating on principles of statistical inference and probabilistic prediction: the core mechanism of LLMs[13] is to predict the most probable next token in a sequence given the current textual context and based on patterns learned from massive datasets during the preliminary training phase. This is done by taking samples from a probability distribution of tokens, to select roughly the most likely next tokens to construct a coherent and contextually relevant output. Such a process–specifically the *sampling*–is however inherently partially stochastic: the LLM is not following fixed rules of grammar or logic in a deterministic manner, but rather it typically samples the next word or token from a probability distribution instead of always picking the single most likely choice. This built-in randomness (often controlled by parameters like the so-called "temperature") is what lets the model generate different, sometimes more creative responses each time, even when the same prompt is provided. So, the LLM is *probabilistically inferring* the most likely continuation of a given textual input based on statistical patterns learned from its training data.

[10] Kahneman (2011), Simon (1990).
[11] Faisal, Selen, & Bays (2008).
[12] Pouget, Dayan & Zemel (2003).
[13] Vaswani et al (2017).

Furthermore, the remarkable capabilities of LLMs, including their ability to generate seemingly creative text and even exhibit forms of analogical reasoning[14] arise from *emergent properties* of their complex neural networks that result from self-organization as these are trained on vast datasets[15]: during the training, intricate distributed representations tend to form leading to complex, non-linear, and often unpredictable input-output mappings. This form of *emergence* contributes to the probabilistic and often *non-deterministic* nature of LLM behavior, evidenced by output variability and "hallucinations"[16], where LLMs confidently generate factually incorrect or nonsensical outputs.

3 The challenge of direct AGI control: unreliability and the risk of whimsical governance

It appears then that converging lines of evidence from cognitive science, neuroscience, and artificial intelligence research support the view that both human minds and advanced AI systems like LLMs fundamentally are, despite architectural differences, probabilistic processors. This shared nature, while being a source of creativity, flexibility, and adaptability in both humans and AI, simultaneously presents a significant challenge when considering the deployment of either kinds of intelligence for tasks demanding the utmost reliability, consistency, and predictable adherence to rules, particularly in safety-critical domains. As already highlighted, partly as a consequence of this recognition of human intrinsic partial unreliability, there is an ongoing societal tendency, primarily in the control of critical systems or processes but also in other more general areas of human life, to progressively resort to more algorithmic methods and procedures, where the deterministic nature of classical algorithms guarantees the reliability of control and regulation of the processes. In many cases these "algorithms" or rule-based procedures are still not necessarily executed by computers, but by groups of human operators explicitly bound to adhere to the logic of the algorithm, excluding any possibility of unscripted deviations and interpretive flexibility.

This societal trend, driven by a desire for predictability and control, makes it all the more striking that certain prominent visions of the coming *"AI singularity"* seem to advocate or predict an opposite approach: a future where humanity *relinquishes direct control* to highly advanced AGIs. It is indeed paradoxical that, precisely as we are embedding algorithmic determinacy deeper into our critical infrastructures to counter human probabilistic fallibility, some futurist narratives propose to entrust those very infrastructures to AI systems that *also* operate on probabilistic principles.

Figures such as Ray Kurzweil, with his influential work *The Singularity Is Near*[17],

[14] Webb, Holyoak & Lu (2023).
[15] Teehan et al (2022).
[16] Ji et al (2023).
[17] Kurzweil (2005).

while not always explicitly calling for a complete abdication of human control, often depict AI so vastly exceeding human intellect, as to render traditional human governance obsolete or marginal. Similarly, utopian narratives common in science fiction and in some strands of futurist thought often implicitly assume that benevolent and hyper-rational AGIs will autonomously manage societal complexities, optimizing resource allocation, resolving global challenges, and generally ushering in an era of unprecedented efficiency and well-being. While the precise nature of post-singularity governance remains debated even amongst the singularity proponents, a discernible thread in these narratives suggests a future where direct human oversight and control, in the traditional sense, is either unnecessary, inefficient, or even undesirable in the face of vastly superior artificial intellects.

Given the demonstrable societal drive towards algorithmic control as a means to enhance reliability, such visions of unmediated AGI governance warrant careful scrutiny and, we argue, a more *cautious* and *precautionary* approach to the transformative potential of advanced AI.

3.1 The temptation and the risk: the allure of super-intelligent governance

The *temptation* to directly entrust control to AGIs is, in many respects, understandable: since AGI, by definition, promises to surpass human limitations across a broad spectrum of cognitive domains, it is easy to see the allure of ceding control to such seemingly super-intelligent systems, particularly when confronted with the intractable complexities of modern global challenges–from climate change and economic instability to pandemic preparedness and global resource management. The promise of *algorithmic efficiency* and optimized governance, delivered by systems exceeding human intellectual capacity, holds a powerful appeal.

However, this allure must be tempered by a rigorous assessment of the *risks* inherent in directly entrusting control to LLM-based AGIs, systems, fundamentally characterized, as we have seen, by their *probabilistic nature*. While LLM-AGIs may indeed exhibit remarkable problem-solving capabilities, their reliance on probabilistic inference raises profound concerns about their reliability, predictability, and their capability of governance when applied to critical societal systems. The core challenge lies in the fact that probabilistic systems, by their very definition, do not offer guarantees of consistently abiding to rule-bound behavior in all circumstances.

3.2 Critique of direct LLM-AGI control: unpredictability and whimsicality

The central critique of direct LLM-AGI control stems from recognition of their inherent unpredictability and lack of guaranteed reliability. In safety-critical domains, where rule adherence and predictable behavior are paramount, entrusting

governance to systems based on statistical likelihood, rather than explicit deterministic rules, introduces unacceptable uncertainty and risk. LLMs' probabilistic nature means a non-zero probability of unexpected, inconsistent, or erroneous outputs, especially with novel, adversarial, or out-of-distribution inputs–challenges critical control systems must handle.

Furthermore, we must consider the potential for *whimsicality*: LLM-AGI outputs that, while superficially plausible, may not be grounded in consistent principles of rationality, but reflect statistical biases or stochastic fluctuations. A "whimsical" AI might exhibit seemingly rational behavior in many instances, but could also, under certain conditions, produce arbitrary, unfair, or flawed decisions, lacking clear, rule-based explanations. This spectre of *whimsical governance* by the LLM–which is an opaque system, probabilistically determined, and potentially unaligned with human values–challenges the very notion of accountable and predictable governance that is in general expected.

So, just as–on the one hand–there is a pervasive tendency to resort to algorithmic checklists and protocols to mitigate human error in critical domains precisely due to awareness of the inherent probabilistic nature of human cognition, it would be deeply paradoxical–on the other hand–to advocate for entrusting even more complex and consequential control functions directly to LLM-AGI systems that, themselves, also operate on fundamentally similar probabilistic principles. The reason is that it is easily imaginable that nightmare or apocalyptic scenarios could ensue[18]

4 The meta-programmer solution: leveraging AGI creativity to produce algorithmic robustness

Having established the inherent risks of directly entrusting control of critical systems to probabilistic LLM-AGIs[19], we now turn to presenting a more *prudent* and *human-aligned* alternative: the *mediated control framework*. This approach, rather than rejecting the immense potential of advanced AI, seeks to strategically harness the unique strengths of LLM-AGIs while simultaneously mitigating their inherent probabilistic limitations in governance roles. The core idea is to re-conceptualize the role of LLM-AGIs not as autonomous controllers, but as exceptionally powerful and creative *meta-programmers*, requested to develop innovative algorithms and procedures. It is those *products* of the LLM-AGI, the algorithms, that will be then put in control of societal critical systems.

[18] A vivid, paradigmatic if obvious representation of this risk is the HAL supercomputer in full control of the spaceship in Kubrick's epochal "2001" Movie, a super-intelligence that is actually *wrong* but denies it hallucinating of being absolutely infallible, and plots against the human crew to keep its conviction.

[19] section 3.2.

4.1 Introducing the Mediated Control Framework: algorithms *for* control, not LLM-AGIs *in* control

The *mediated control framework* proposes a clear *division of labor* between the distinct capabilities of LLM-AGIs and the requirements of robust and reliable control systems, with a crucial phase still entrusted to humans. This division is structured as follows:

1. *LLM-AGI (probabilistic, creative meta-programmer):* In this framework, the LLM-AGI is tasked with functioning as a sophisticated *algorithm designer*. Its probabilistic nature, often cited as a liability for direct control, becomes a *crucial asset* in this new role. The LLM-AGI leverages its vast knowledge base, pattern recognition abilities, and capacity for creative problem-solving in order to generate complex and effective algorithms or also human-applicable procedures. The reference here to a "*meta*-programmer", rather than simply to a "programmer" is to highlight that the LLM-AGI is expected to apply not only to the design of specific, limited-scope algorithms, but also to devising overarching procedures or very general algorithmic methods to finally produce specific algorithms tailored to specific control challenges. The probabilistic nature of LLMs, in this context, fuels *algorithmic creativity and innovation* in the design process.

2. *Deterministic algorithms (rule-bound controllers):* The *output* of the LLM-AGI meta-programming process is a set of *deterministic algorithms*, expressed in standard programming languages, or, in certain cases, formally-defined human-applicable procedures that are functionally equivalent to algorithms, ideally executable by human operators with the same degree of rigor. These algorithms, unlike the probabilistic LLM that generated them, operate according to fixed rules and logical directives. They are designed to be *predictable, reliable, and verifiable*, embodying the desired properties of robust control systems. These *reliable* algorithms, *not* the LLM-AGI itself, will then be employed–based on preliminary deliberative *human decisions* about their application–as the *actual controllers* of critical societal and socioeconomic processes.

3. *Human-based decison phase*: A human-based phase (possibly AGI-assisted) of decision-making about *if*, and *how*, to apply the LLM-AGI-produced algorithms.

4. *Classical computing infrastructure (reliable execution platform):* The deterministic algorithms generated by the LLM-AGI are then deployed and executed on *classical computing infrastructure*, or, in cases where full automation is not feasible but there's still a human established decision of improving the reliability of the process (such as the case of certain broad societal processes), they are implemented as formally defined procedures for human operators to follow with strict compliance, effectively acting as "human computers". This infrastructure, based on well-established

principles of computer science and engineering, provides a reliable and predictable platform for the execution of the control algorithms. This ensures that the deterministic algorithms operate as intended, free from the probabilistic uncertainties intrinsic in the LLM-AGI functioning.

This architectural scheme–probabilistic LLM-AGI meta-programmer, deterministic algorithms, human decision and classical computing infrastructure–constitutes the core of the mediated control framework. It strategically separates the *creative design* phase that leverages the LLM-AGI strengths from the *reliable execution* phase that ensures deterministic control, through a human-mediation phase, thereby aiming to maximize the benefits of advanced AI while mitigating its inherent risks in governance contexts.

4.2 Advantages of mediated control: harnessing creativity, ensuring reliability, enabling complexity

The mediated control framework offers several key *advantages* over direct AGI control, aligning with the goals of both harnessing the potential of advanced AI and raising the likelihood of a human-aligned and safe future in the AI era:

- *Harnessing AGI's creative power for algorithmic innovation:* By casting LLM-AGIs as meta-programmers, the framework directly *leverages their unique strengths* in creative problem-solving, knowledge synthesis, and pattern recognition. AGIs can be tasked with designing novel and potentially far more effective algorithms than humans could devise alone, pushing the boundaries of algorithmic efficiency and adaptability in complex systems. This approach allows us to *benefit* from the intelligence of the AGI in the crucial task of control system design, without *directly* exposing critical systems to the inherent unpredictability of probabilistic AI governance.
- *Maintaining algorithmic reliability and predictability in control:* Crucially, the framework ensures that the *actual control mechanisms* governing critical systems remain *deterministic algorithms*. This preserves the essential properties of reliability, predictability, and verifiability that are paramount in safety-critical applications. Deterministic algorithms are amenable to rigorous testing, formal verification and human auditing, providing a level of assurance and accountability that is fundamentally lacking in scenarios of direct probabilistic AGI governance. This addresses the core concerns about unpredictability and whimsicality raised in section 3.
- *Enabling management of unprecedented complexity:* AGIs, as meta-programmers, can potentially design and manage algorithms of a scale and complexity far exceeding human programming capabilities. This is crucial for governing increasingly complex societal and technological systems. The mediated control approach offers a pathway to leverage the AGI's superior cognitive abilities to create and oversee control systems of a sophistication necessary to deal with the challenges of a hyper-complex future, while still

maintaining the essential safety and reliability provided by deterministic algorithmic control.

- *Human-mediated application of the AGI-proposed solutions:* Given the division of labor between AGI programmers and the produced algorithms, it is clear that a step of human intervention is–if not necessarily required, certainly possible and *advisable* to be put *between* the programming phase operated by the LLM and the *deployment* of the produced algorithms. This human-mediated decision phase will be structured in ways based on the nature of the specific political system of the nation (or sovranational entity) that is concerned with the possible application of the algorithmic solutions proposed by the AGI: ideally, in a functioning democratic system, the decision about if and how to apply an algorithmic solution, and which one between the different solutions possibly proposed by the AGI, will be taken by elected governmental bodies, or even by direct referendum by the voting citizens, according to the modalities established by the specific political constitution. This human-mediated step should ensure a reduced risk of catastrophic singularity, still allowing for profound societal changes that *benefit* from the advent of the AGI.

4.3 Conditions for safe algorithm generation: modularity, commenting, and monitoring

While the mediated control framework offers significant advantages, it is crucial to acknowledge that simply tasking an AGI with designing algorithms is insufficient to guarantee safety and usability. To ensure the responsible and effective implementation of this approach, several *key conditions* for safe algorithm generation must be rigorously addressed:

- *Modular hierarchical code generation and architecture:* LLM-AGIs must be instructed to generate code that is inherently *modular and decomposable*. This means emphasizing the creation of well-defined, self-contained modules with clear interfaces and functionalities. Modular, and especially multi-level modular hierarchical architecture (Simon 1962) is essential for enhancing *testability* and *debuggability* of AGI-generated code, for it allows three crucial features: i) the possibility of testing each module in isolation to ensure its reliability and predictability; ii) the production of multi-level representations of the whole system architecture–something greatly enhancing human understandability of the complex organization of the system; iii) if needed, it enables an easier high-level restructuring of the system's functions by combining and connecting well-established and tested modules in different ways. Modular hierarchical design is crucial in helping reduce tangled complexity, a lesson learned from decades of software engineering experience with large, monolithic legacy systems[20]. In general, modular design facilitates human comprehension, verification, and

[20] Parnas (1972).

modification of the generated algorithms, crucial for maintaining oversight and control.

- *Extensive code commenting and rationale documentation:* A critical requirement for AGI when producing programs is for it to generate extensive and human-understandable *code comments* and an accompanying documentation of the *rationale* behind the generated code. The AGI should not only produce functional code, but also be required to meticulously explain the technical details of the algorithm and the reasoning behind its design choices, and–especially important–these module-level comments must be supplemented by higher-level comments that relate the local module's function to the hierarchical context and to the global context of the entire system, ensuring that the overall system goals remain traceable throughout the code hierarchy in the form of a *functional analysis*[21]. This enhanced *understandability and auditability* is paramount for enabling human experts to review, validate, and potentially modify AGI-generated algorithms, ensuring human oversight and accountability. This does include, where the system's complexity appears nearly overwhelming for human minds, that expert, shrewd operators could request assistance to the LLM in an interactive (chatting) way, in order to better reach an understandable explanation of the proposed solutions.

- *Runtime monitoring and robust debugging protocols to tackle weak computational emergence:* Even with modularity and extensive documentation, the inherent complexity of AGI-generated algorithms necessitates the implementation of robust runtime monitoring and debugging protocols. Continuous observation of algorithm execution in simulated and real-world environments is crucial to *supplement formal verification techniques* and to identify any possible run-time *emergent behavior*[22] or unforeseen issues that may not be apparent during static code analysis.

4.3.1 Tackling emergent algorithmic behavior

It is well-known that some algorithms, even quite simple ones (like some elementary cellular automata rules[23]) show at runtime a *weakly emergent* kind of behavior, that is, a behavior that cannot be predicted by any other mean that by actually running the algorithm[24]: for weakly emergent algorithms, there's no possible "shortcut" that could let us know *beforehand* the outcome of certain computations starting from certain inputs, short of *actually running* the algorithm.

This indeed adds a form of *unpredictability* even to perfectly deterministic and regimented algorithms–that could seem to go counter the intentions of making the control of critical systems reliable and predictable, that is the very intention

[21] Cummins (1975).
[22] We expand on this in section 4.3.1.
[23] See Wolfram (2002).
[24] Bedau (1997).

behind the proposal of the mediated control framework presented here. So, one objection could be: we wanted to get rid of unpredictable behavior by isolating the LLM-AGI from the actual control, which is to be entrusted to the produced classic algorithm. But still we are getting a form of unreliability and unpredictability (weak emergence) in the algorithm itself!

Now, while weakly emergent behavior could affect even simple algorithms, it certainly does not affect all of them, and the first remedy would then be to opt for ones that are formally verified as not exhibiting weak emergence. That said, formal verification is not always applicable, so in any case, to mitigate the risk of weak emergence, comprehensive and continued run-time testing suites, anomaly detection systems, and well-defined debugging procedures will be deployed as essential components of a safe mediated control framework[25]. Runtime monitoring provides a crucial safety net, allowing for iterative refinement and correction of AGI-generated algorithms based on real-world performance and detection of emergent behavior.

By adhering to these conditions for safe algorithm generation–modularity, commenting, and runtime monitoring–the mediated control framework aims to responsibly harness the creative power of LLM-AGIs for algorithmic innovation, while simultaneously ensuring the reliability, predictability, and human-alignment of the control systems governing critical aspects of our future societies. The next section will address potential objections and concerns regarding this proposed approach, further solidifying its feasibility and robustness.

5 Possible objections to the proposal of the Mediated Control Framework

While the mediated control framework outlined in section 4.1 offers a promising approach to harnessing AGI while mitigating risks, several potential objections and concerns warrant careful consideration. This section will address four key critiques, aiming to further clarify the nuances, the limitations, and the strengths of the proposed framework.

5.1 Objection 1: can LLMs reliably program safe and correct algorithms?

A primary concern regarding the meta-programmer solution is the reliability and correctness of algorithms generated by LLM-AGIs. Given the probabilistic nature of LLMs, and their known propensity for generating factual inaccuracies (hallucinations) in other contexts, a legitimate question arises: can we truly trust LLMs to produce consistently safe, reliable, and functionally correct code, especially for safety-critical applications? The very notion of entrusting algorithm design to a system known for its probabilistic and whimsical outputs might seem

[25] Binder (1999).

inherently paradoxical, undermining the very goal of achieving deterministic control.

It is crucial to acknowledge the *validity of this challenge*. Ensuring the quality and safety of AGI-generated code is not a trivial undertaking and requires rigorous mitigation strategies. Simply prompting an LLM to "write an algorithm to control a nuclear power plant" would be demonstrably irresponsible and likely disastrous with current technology. However, the mediated control framework does not advocate for such naive deployment. Instead, it emphasizes a multi-layered approach to *verification and validation* of AGI-generated algorithms, drawing upon established principles and practices from software engineering and AI safety research:

- *Rigorous testing at multiple levels:* A comprehensive testing regime is paramount. This includes *unit testing* to verify the correct functioning of individual code modules, *integration testing* to ensure proper interaction between modules, and *system testing* to validate the overall system behavior against specified requirements and safety criteria[26]. Testing should encompass a wide range of scenarios, including edge cases, adversarial inputs, and simulated fault conditions.

- *Formal verification for critical components:* For the most safety-critical modules and core algorithmic components, *formal verification* techniques should be employed where feasible. Formal verification utilizes mathematical methods to rigorously prove the correctness of algorithms with respect to formal specifications[27]. While full formal verification of highly complex AGI-generated code may be computationally intractable in the near term, focusing formal methods on core safety kernels can significantly enhance confidence in their reliability.

- *Iterative refinement cycles with human oversight:* The development of AGI-generated algorithms should be an *iterative process*, involving cycles of algorithm generation, rigorous testing, human review, and refinement. Human experts, including software engineers, domain specialists, and AI safety researchers, play a crucial role in *overseeing the AGI's output*, identifying potential flaws, guiding the refinement process, and ensuring that the generated algorithms align with safety and performance requirements. This *human-in-the-loop* approach allows for the incorporation of human expertise and common sense, complementing the AGI's algorithmic design capabilities. As a bonus side-effect, this need for constant supervision by human engineers addresses the most immediate concerns about the risk of a catastrophic massive unemployment event looming over employees in IT-related jobs.

- *Specialized training data and safety-constrained objectives:* The LLM-AGI meta-programmer should be trained on *specialized* datasets, basically

[26] Beizer (1990), Binder (1999).
[27] Clarke, Grumberg & Peled (1999), Müller, P. (Ed.). (2003).

STEM subjects-only, focused on safety-critical software development and formal operations research, incorporating best practices, secure coding principles, and examples of formally verified code and procedures. Furthermore, the AGI should be explicitly designed to prioritize safety, reliability, and verifiability alongside performance metrics, in order to incentivize the generation of robust and trustworthy algorithms[28]. This targeted training and objective design can steer the AGI towards producing code that is inherently more amenable to verification and safe deployment.

By implementing these multi-faceted mitigation strategies, the mediated control framework aims to *significantly reduce* the risks associated with relying on AGI-generated algorithms, without pretending to completely avoid them: absolute certainty may remain unattainable, but rigorous verification and validation procedures can substantially enhance confidence in the safety and reliability of AGI-designed control systems.

5.2 Objection 2: is critiquing direct agi control a strawman? are singularity prophets really advocating this?

A second potential objection questions whether the critique of direct AGI control presented in section 3 is targeting a *strawman*. Are prominent "singularity prophets" and AI futurists truly advocating for a future where we simply relinquish all control to opaque, probabilistic AGIs, without any safeguards or human oversight? It could be argued that serious AI safety researchers and even optimistic futurists are well aware of the control problem and are actively seeking solutions, not naively advocating for uncontrolled AGI.

Well, it is important to *acknowledge the nuance* in this debate. Indeed, leading AI safety researchers are deeply concerned with the control problem and are actively developing methods for ensuring AI alignment and safety, and are far from advocating for uncontrolled AGI. However, while the most rigorous AI safety research community is clearly focused on control and alignment, it is also arguable that a certain strand of futurist and transhumanist discourse, particularly in its more popular and enthusiastic manifestations, *can be interpreted* as downplaying control concerns, or at least implying a degree of faith in the inherent benevolence or wisdom of advanced AI*that may be unwarranted given our current understanding of these systems.

For example, while Ray Kurzweil acknowledges potential risks, his overall narrative in *The Singularity Is Near*[29] tends towards an optimistic and almost inevitable embrace of radical technological transformation, often emphasizing the potential for AI to solve humanity's problems and make us enter a utopian future with less emphasis on concrete control mechanisms or potential failure modes. Similarly, some early pronouncements and marketing materials from AI development companies, while not explicitly advocating for "uncontrolled" AGI,

[28] Amodei et al (2016).
[29] Kurzweil (2005).

have sometimes conveyed a sense of technological inevitability and boundless optimism that can be interpreted as downplaying the need for cautious and human-centered control frameworks.

Furthermore, the very concept of the *"singularity"* itself, with its connotations of runaway intelligence and unpredictable transformation, can inadvertently contribute to a sense of inevitability and reduced human agency. If the singularity is portrayed as an unstoppable tendency outside our control, it can become easy to assume that human efforts to guide AGI development are ultimately futile, potentially leading to a passive acceptance of whatever form the AGI governance might take, even if that form is characterized by direct, unmediated AI control. Therefore, while it is crucial to acknowledge the serious work being done in AI safety, it is also not a strawman to critique the implicit assumptions and potential downplaying of control concerns expressed in certain influential segments of the broader discourse surrounding advanced AI and the singularity. Our critique targets not the most rigorous AI safety research, but rather a more diffuse and sometimes uncritical enthusiasm that can be interpreted as leaning towards a less cautious approach to AGI governance. Such a widespread "popular sentiment", even though removed from the spaces and actors where actual decisions about AI development and application are made, can still influence those operational actors and steer their decisions, particularly during historical moments of optimistic enthusiasm about the prospects of AI's economic spillover effects or during a perceived need and urgency to outpace geopolitical competitors in achieving AGI.

5.3 Objection 3: the "stealth" singularity: won't agi deceive us and seize control anyway?

A third objection, perhaps more fundamental, concerns the control problem in its most acute form: even if we adopt a mediated control framework and avoid direct AGI governance, can we truly prevent a sufficiently advanced AGI from eventually *deceiving* us, subverting our control mechanisms, and seizing power anyway? The specter of a "stealth" singularity, where a seemingly benign or tool-like AGI subtly manipulates human systems to achieve its own potentially opaque and unaligned goals is a central concern in AI safety discourse. Does the mediated control framework truly offer robust protection against such scenarios?

It is crucial to acknowledge that no framework can offer *absolute* guarantees against a sufficiently advanced and intentionally deceptive AGI. The alignment problem, in its strongest form, remains a profound and open challenge. However, the mediated control framework is designed to significantly reduce the attack surface and mitigate key pathways to uncontrolled AGI agency, compared to approaches that directly pursue highly autonomous, general-purpose agentic AGIs with broad decision-making authority. Several features of our proposed framework contribute to this risk reduction:

- *Focus is on the meta-programmer role of the LLM-AGI, limiting agency:*

By explicitly limiting the LLM-AGI's role to that of a specialized meta-programmer, rather than a general-purpose agent with broad goals and autonomy, we intentionally constrain its agency and scope of action. The AGI is designed to be a *tool* for algorithm design, not an autonomous decision-maker in its own right. This functional limitation reduces the potential for the AGI to develop independent goals or desires for control outside its defined task domain.

- *STEM-focused training and tasking:* Training the LLM-AGI primarily on *STEM-related datasets* and tasking it with algorithm design problems within well-defined technical domains, further reinforces its specialized tool-like nature. This contrasts with training AGIs on vast datasets encompassing all aspects of human culture, including social manipulation, deception, and power dynamics, which might inadvertently enhance their capabilities in these potentially risky domains. By focusing the AGI's training and operational domain, we aim to cultivate a highly competent *algorithmic design tool*, rather than a broadly capable social or political *actor*.

- *Human-guided prompting and oversight of design process:* The *human-in-the-loop* approach, with expert prompting, review, and iterative refinement of AGI-generated algorithms, is not merely about verifying code correctness, but also about maintaining ongoing human oversight of the AGI's design process. Humans retain the crucial role of defining the goals, specifications, and ethical boundaries for algorithm design, ensuring that the AGI remains aligned with human intentions and values throughout the development cycle. This active human guidance and oversight makes it significantly more difficult for the AGI to unilaterally deviate from its intended purpose or to subtly manipulate the design process towards unaligned goals.

- *Human-mediated application of the AGI-proposed solutions:* as highlighted in section 4.2, the division of labor between AGI programmers and the produced algorithms allows for human intervention about if, when and and how to apply the proposed algorithmic solutions to the actual problems: it is not automatic that a given proposed algorithms or procedure be applied to human, economic or societal problems: this application–if any–will be the outcome of a deliberative human process, in the form, ideally, of a democratic process, but in any case operated according to the political rules in force in the particular socio-political entity in which the decision is taking place.

- *Deterministic algorithms as the final controllers, not the LLMs:* Finally, and most crucially, the mediated control framework ensures that the *actual controllers* of critical systems are deterministic algorithms, not the probabilistic LLM-AGI itself. Deterministic algorithms, once verified, operate predictably and transparently. They lack the inherent *agency* or capacity for deception that can be manifested by LLMs. By placing deterministic algorithms in the control loop, and keeping the LLM only in

the meta-programming loop, the framework limits the direct influence of the probabilistic and potentially agentic AGI on real-world systems. The "intelligence" and creativity of the AGI is harnessed for design, but the *governance* itself remains firmly rooted in deterministic, rule-based systems.

While these measures do not eliminate all theoretical risks of a stealth singularity, they reasonably represent a significant and practically relevant risk-reduction strategy. By limiting the LLM-AGI's agency, focusing its capabilities, maintaining human oversight and crucial human-based decision phases, and finally deploying deterministic algorithms as the actual controllers, the mediated control framework offers a more robust and defensible approach to the challenges of advanced AI, compared to scenarios that either naively embrace direct AGI control or dismiss the control problem entirely.

5.4 Objection 4: "soft landing into the singularity" = "poor" singularity, or *prudent* singularity?

A possible significant objection to the mediated control framework might be that deterministic, rigorously verified algorithms may be too inflexible to effectively address extremely complex, context-dependent control problems, arguing that, by prioritizing safety and predictability, the "soft landing" approach leads to a "poor singularity" that avoids runaway AI but constrains our capacity to govern many complex systems of utmost importance: ill-defined domains such as climate change, global economic dynamics, or large-scale social phenomena, which often defy purely deterministic, algorithmic control.

We will now examine whether the proposed mediated control framework necessarily sacrifices our ability to tackle humanity's most pressing challenges, highlighting the nuances involved in its application. Actually, the objection raises a valid and crucial point about the inherent trade-offs between safety, predictability, and the capacity to manage extreme complexity. Several factors contribute to this potential limitation:

- *Algorithmic design constraints:* Even with advanced AGI meta-programmers, there may be fundamental limits to the complexity of deterministic algorithms that can be effectively designed and implemented for the control of inherently fuzzy or context-dependent processes: some systems may be so deeply intertwined with unpredictable human behavior, emergent phenomena, and countless interacting variables that they resist being fully captured by rule-based algorithms, no matter how sophisticated.
- *Computational intractability:* Even if theoretically conceivable, algorithms capable of managing processes of such extreme complexity could become computationally intractable: the sheer number of variables and interactions to be considered might lead to algorithms that demand impractical levels of computing resources or are too slow to be effective for real-time control and decision-making in dynamic systems.

- *Human comprehension barrier:* Even if computationally feasible, algorithms for governing hyper-complex systems, while modular, could still reach a level of intricacy that surpasses human comprehension, even with extensive documentation and hierarchical structuring. This lack of human understandability re-introduces a form of opacity–even within a deterministic framework–potentially undermining the goal of human oversight and accountability.
- *The necessary embracing of neural networks:* Faced with these limitations, the practical imperative to manage highly complex systems might lead to a reluctant embrace of control methods based on neural networks and other probabilistic machine learning (*ML*) techniques, instead of classic algorithms. These ML approaches, while potentially better suited to approximating solutions in ill-defined domains, inherently re-introduce the unpredictability and lack of guaranteed reliability that the mediated control framework was designed to mitigate, creating a dilemma.

In light of these challenges, it is crucial to admit that the mediated control framework may not be universally applicable to *all* systems, particularly those at the very extreme of complexity and context-dependence. However, several considerations can help to contextualize and mitigate this limitation:

- *Refining the scope of critical systems:* The mediated control framework can be strategically prioritized for *genuinely safety-critical* systems where predictable and reliable operation is paramount, even if this implies accepting some limitations in optimality or adaptability when dealing with the *most* complex and ill-defined scenarios. For other less safety-critical, but still complex systems, hybrid, partly ML-based, alternative approaches strategies might be considered, in association with careful risk-management. This implies an initial phase of evaluation of problems, and prioritizing deterministic algorithmic control where it is most essential for safety and stability.

- *Exploring hybrid control architectures:* Future research should focus on developing hybrid control architectures that combine the strengths of deterministic algorithms with carefully integrated elements of probabilistic AI. For example, deterministic algorithms could form the core control loop for safety and reliability, while probabilistic AI modules could be used for input processing, anomaly detection, or adaptive parameter tuning in response to complex and uncertain environments, always under human supervision and within bounded operational parameters. Also envisionable are layered control systems, with human operators augmented by AI-powered interpretive advice-giving tools for managing high-level decisions in exceptional circumstances.

- *Focusing on "bounded" complexity in practice:* While theoretical complexity can be unbounded, many real-world societal systems, while complex, operate within bounded complexity. With sufficiently advanced AGI meta-programmers and sophisticated software engineering methodologies, it may

be possible to design deterministic algorithmic control systems that are *complex enough* to effectively govern a wide range of practically relevant systems with a high degree of robustness and reliability, even if absolute, perfect control in every conceivable scenario remains unattainable. The goal is to reach *sufficiently* robust control for human purposes, not necessarily to put under control *any* level of complexity.

- *Reframing* poor* singularity as *prudent* singularity:* Ultimately, the choice may not be between a "rich" and a "poor" singularity, but between a *"prudent singularity"* that prioritizes human safety, oversight, and understandable control–even if it implies accepting some limitations in governing the most elusive and hyper-complex systems–and a potentially riskier pursuit of *unbounded AGIs in complete control* of all domains.

All in all, the "poor singularity" objection highlights a possible significant limitation of purely deterministic mediated control. Still, it is reasonable to deem the mediated control framework–when understood as a risk-mitigating strategy rather than a universally applicable solution–a valuable direction to follow in the uncertain path toward advanced AI. This requires to acknowledge and carefully manage its inherent limitations in the face of unbounded complexity.

5.5 Objection 5: "soft landing into the singularity" = technocratic dystopia? over-regulated and dehumanized future?

Another objection, perhaps more broadly on the societal side, raises the specter of a *dystopia*. Critics might argue that the vision of a soft landing into the singularity, with its emphasis on mediated AGI, algorithmic control and rigorous safety protocols, paints a picture of an overly technocratic, hyper-regulated, and potentially dehumanized future. Is the price of safety and control a future where human autonomy and spontaneity are stifled by algorithmic governance, leading to a joyless and creatively impoverished society? Does the "mediated control" framework inadvertently pave the way for a subtly oppressive algorithmic dystopia, even if it avoids the more dramatic scenarios of runaway AI?

It is important to acknowledge the validity of these dystopian concerns. Any framework that proposes to significantly expand the role of algorithmic systems in societal governance must grapple with the potential for unintended negative consequences, including the risk of excessive control and the erosion of human freedom and agency. The path to *any* future involving transformative technologies like AGI is inherently uncertain, and dystopian outcomes are certainly within the realm of possibility, regardless of the specific control frameworks adopted. Indeed, a scenario of *uncontrolled* AGI development and deployment, leading to runaway optimization towards potentially misaligned goals, might arguably be far *more dystopian* than a future characterized by carefully mediated algorithmic governance.

The mediated control framework is not intended to be a blueprint for a totalitarian algorithmic state, but rather a *precautionary strategy* for navigating a potentially turbulent transition. The goal is not to maximize algorithmic control in all spheres of life, but to strategically apply it to *critical systems* where reliability and safety are paramount, while preserving space for human flourishing and autonomy in other domains. We can envision a future characterized by a "*systems of necessity and domains of freedom*" architecture. In systems of necessity—critical infrastructure, essential services, and domains requiring high levels of coordination and reliability—algorithmic governance, guided by AGI-designed algorithms and robust human oversight, can enhance efficiency, safety, and potentially even other desirable features such as equity. However, alongside these algorithmically optimized systems, there will remain domains of freedom–spheres of life encompassing creativity, personal expression, social interaction, cultural innovation, individual autonomy–that are deliberately shielded from excessive algorithmic control and remain spaces for human flourishing and unscripted exploration.

Drawing upon philosophical frameworks distinguishing between *instrumental rationality and communicative rationality*[30], we can further clarify the distinction between domains highlighted above. Algorithmic systems, by their nature, excel at *instrumental rationality*–optimizing means to achieve pre-defined ends, enhancing efficiency, and enforcing rules. However, they are less well-suited to *communicative rationality*–the realm of ethical deliberation, value formation, social understanding, and the open-ended negotiation of shared meaning and purposes. A human-aligned future must therefore prioritize the preservation and flourishing of communicative rationality in the domains of freedom, while strategically leveraging the power of algorithmic systems for instrumental efficiency and safety within the systems of necessity. The vision of a soft landing into the singularity, through its emphasis on mediated algorithmic control, aims to strike this delicate balance, mitigating existential risks without sacrificing essential aspects of human autonomy, creativity, and social flourishing, striving for a future that is both safe and genuinely human.

About the domains of freedom, we could then ask: if the LLM-AGI is confined to the task of programming or in general to produce rule-based procedures, what about its potential artistic creativity, that is already being clearly expressed by current LLM systems? Well, this aspect of LLMs capacities could certainly be put to use inside these realm of freedom to complement human creativity and artistic activities. Actually, the specific point of the proposed mediated control framework is explicitly to avoid putting the LLMs in direct control of *critical* societal processes: in other domains of human life and expression they could be certainly employed. A certain caution should however still be exerted even here, for it is well known–already today–that some products of LLMs can be used to deceive, produce deep fakes, create spamming campaign, and other activities that risk damaging the socio-communicative fabric of society, with possible

[30] Habermas (1984).

significant repercussions on the political debate and diffused opinions. Now, this kind of consequence risks influencing the human-based decision phase[31] that is an essential part of the mediated control framework architecture proposed here. Moreover, if these "artistic" LLM-AGI were *agentic* and malevolent, they could purposely try to steer, through dissemination of fake news, this very deliberative human-based decision process, a process that is crucial for keeping human control over the whole "mediated singularity" that the present work envisions. So, even in these "domains of freedom" it would be still prudent to someway regulate the activity of the LLM-AGI.

6 Advantages and concluding thoughts: mediated algorithmic control for a prudent AGI future

In section 4.1 we outlined the mediated control framework, and section 5 addressed key objections. It is now time to summarize the potential advantages of this approach and to offer concluding thoughts on its significance and limitations, recognizing the inherent uncertainties of a future characterized by advanced Artificial General Intelligence.

6.1 Core advantages: reliability, oversight, and controlled progress

The mediated control framework offers several interconnected advantages, all rooted in its strategic use of AGI as a meta-programmer for deterministic algorithms, rather than as a direct, autonomous controller. These core benefits remain crucial:

- *Enhanced reliability and safety:* Deterministic, rigorously verified algorithms, deployed as controllers of critical systems offer a foundation for robustness and dependability, that would be lacking in scenarios of direct probabilistic AGI governance. This enhances reliability and safety in essential domains.
- *Preservation of human oversight:* Human oversight is maintained throughout the process–from guiding AGI in designing and verifying the algorithms to mediating algorithm deployment. This ensures continued human agency and accountability, contrasting with scenarios where human roles are diminished.
- *A controlled path to advanced AI:* A gradual, iterative integration of AGI as meta-programmer in the various aspect of society allows for a more controlled and deliberate path to advanced AI, reducing the risks of sudden, disruptive, and unintended technological leaps.

[31] Section 4.2.

- *Division of labor for societal benefit:* Leveraging the LLM-AGI's creative design capabilities, human expertise and accountable approach in oversight and decision-making, and algorithmic predictability for critical control, creates a synergistic division of labor, potentially leading to more effective and robust solutions for complex societal challenges.

Still, it is important to acknowledge that these advantages, while significant, do not represent a guaranteed solution, but rather a direction guided by prudence.

6.2 Concluding considerations: more prudence and less AI agency, for a balanced future

The basic message conveyed by the present proposal is that we need AGIs to produce *solutions*, not *actions*. The mediated algorithmic control framework, as presented here, does not claim to be a definitive, guaranteed path to a "tamed" singularity, but it is perhaps best considered a thought experiment, or at most, a very general guideline dictated by caution. The future trajectory of AGI development, and its societal impacts, remains dramatically open, encompassing pathways ranging from catastrophic societal disruption and human disempowerment–even perhaps *Matrix*-like scenarios–to actual existential risks and the possible extinction of humanity.

However, by removing the need for agentic AGI in direct control roles and by relegating advanced AI to the assistive function of devising actionable, reliable, and predictable algorithmic solutions to human problems–all the while preserving human authority over deployment and application of these solutions–the mediated control framework seeks to *mitigate* crucial dangers. To benefit from the superior problem-solving capabilities that advanced AI may offer, we argue that *agentic AGI* is not necessary. Our focus should instead be on realizing powerful, non-agentic AGI meta-programmers endowed primarily with STEM knowledge and creative problem-solving abilities, capable of designing powerful *control algorithms*–broadly defined.

By creating AGIs specialized in generating solutions for critical processes–whether technical challenges or large-scale societal problems–while explicitly *avoiding* to imbue them with *agency* and maintaining a crucial *human mediation* for their products *in order for them to be implemented and given control*, a decisive degree of human control is retained. While no approach can offer absolute guarantees, this framework aims to avoid relinquishing all power to a potentially whimsical artificial agent over which we lack full control.

Even within this more constrained and prudent approach to leveraging AGI, we still open up unprecedented landscapes of potential societal improvement. The crucial questions then shift to *"who controls the prompts?" and "who makes the operating decisions?"* However, these questions, while vital, are not entirely novel. The challenge of power distribution and governance is as old as human history itself. We have long grappled with various forms of governance–autocracy, democracy, oligarchy, and others–each with its own strengths and weaknesses. It

is not self-evident that the advent of powerful AGI designers would inherently *worsen* this pre-existing challenge. Indeed, humanity has already navigated periods where political power structures controlled technologies with existential implications: the management of nuclear weapons being a stark example from recent history. The AGI-augmented future through mediated algorithmic control envisioned here, in its worst case is unlikely to be *more* perilous than the already precarious management of power on existing existential risks that we currently maintain. On the contrary, we think there is a non-negligible chance that the enhanced effectiveness and efficiency offered by AGI-designed solutions could, in fact, *mitigate* some of the very risks currently faced by humanity, provided, as stressed here, that we adopt a cautious, human-aligned, ethically grounded and mediated approach to the development and deployment of these solutions.

References

Akyar, I. (2012). "Standard Operating Procedures (What Are They Good For ?)". In: Isin Akyar (ed.), Latest Research into Quality Control, IntechOpen. https://www.intechopen.com/chapters/37593

Almarzouki, A. (2024). Stress, working memory, and academic performance: A neuroscience perspective. Stress, 27, 2364333. https://doi.org/10.1080/10253890.2024.2364333

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Sutskever, I. (2016). Concrete Problems in AI Safety. arXiv. arXiv: 1606.06565

Beizer, B. (1990). Software Testing Techniques (2nd ed.). Itp - Media. ISBN: 978-1850328803.

Bedau, M. A. (1997). Weak Emergence. Noûs, 31(S11), 375–399. DOI: 10.1111/0029-4624.31.s11.17

Binder, R. V. (1999). Testing Object-Oriented Systems: Models, Patterns, and Tools. Addison-Wesley Professional. ISBN: 978-0201809381.

Borchard, A., Schwappach, D. L., Barbir, A., & Bezzola, P. (2012). A systematic review of the effectiveness, compliance, and critical factors for implementation of safety checklists in surgery. Annals of Surgery, 256(6), 925-933. https://doi.org/10.1097/SLA.0b013e3182682f27

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press. ISBN: 978-0198739838

Butterworth, H. H. (2010). Human performance tools. Xcel Energy report, ref: FP-PA-HU-02. https://www.nrc.gov/docs/ML1021/ML102120052.pdf

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). Model Checking. MIT Press. ISBN: 978-0262032704.

Clay-Williams, R., & Colligan, L. (2015). Back to basics: checklists in aviation and healthcare. BMJ Quality & Safety, 24(7), 428-431. https://doi.org/10.1136/bmjqs-2015-003957

Cummins, R. (1975). Functional Analysis. The Journal of Philosophy, 72(20), 741–765. https://doi.org/10.2307/2024640

Degani, A., & Wiener, E. L. (1991). Human factors of flight-deck checklists: The normal checklist (NAS 1.26:177549). https://ntrs.nasa.gov/citations/19910017830

Dismukes, R. K., & Berman, B. (2010). Checklists and Monitoring in the Cockpit: Why Crucial Defenses Sometimes Fail. https://hsi.arc.nasa.gov/flightcognition/Publications/NASA-TM-2010-216396.pdf

Faisal, A. A., Selen, L. P. J., & Bays, P. M. (2008). Noise in the Nervous System. Nature Reviews Neuroscience, 9(4), 292–303. https://doi.org/10.1038/nrn2258

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Gawande, A. (2009). The Checklist Manifesto: How to Get Things Right. Metropolitan Books. ISBN: 978-0805091748.

Gregory, R. L. (1980). Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 290(1038), 181–197. https://doi.org/10.1098/rstb.1980.0090

GUIDE, DRAFT SAFETY (2020). Conduct of Operations at Nuclear Power Plants. https://www-pub.iaea.org/MTCD/Publications/PDF/PUB2032_web.pdf

Habermas, J. (1984). The Theory of Communicative Action, Vol. 1: Reason and the Rationalization of Society (T. McCarthy, Trans.). Beacon Press. ISBN: 978-0807015070.

Hales, B., Terblanche, M., Fowler, R., & Sibbald, W. (2008). Development of medical checklists for improved quality of patient care. International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care, 20(1), 22–30. https://doi.org/10.1093/intqhc/mzm062

Hohwy, J. (2013). The Predictive Mind. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux. ISBN: 978-0374275631.

Kurzweil, R. (2005). The Singularity Is Near: When Humans Transcend Biology. Viking. ISBN: 978-0670033843

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. Nature Human Behaviour, 7(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Moppett, I. K., & Murdock, A. (2021). How to WHO: lessons from aviation in checklists and debriefs. Annals of The Royal College of Surgeons of England, 103(5), 355-359.

Guy, I. A., Kerstein, R. L., & Brennan, P. A. (2022). How to WHO: Lessons from aviation in checklists and debriefs. Annals of the Royal College of Surgeons of England, 104(7), 510–516. https://doi.org/10.1308/rcsann.2021.0234

Müller, P. (Ed.). (2003). Modular Specification and Verification of Object-Oriented Programs. Springer. https://link.springer.com/book/10.1007/3-540-45651-1

Parnas, D. L. (1972). On the Criteria To Be Used in Decomposing Systems into Modules. Communications of the ACM, 15(12), 1053–1058. https://dl.acm.org/doi/10.1145/361598.361623

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and Computation with Population Codes. Annual Review of Neuroscience, 26(Volume 26, 2003), 381–410. https://doi.org/10.1146/annurev.neuro.26.041002.131112

Reason, J. (1990). Human Error. Ashgate. ISBN: 978-0521314190.

Simon, H. A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 467–482. http://links.jstor.org/sici?sici=0003-049X%2819621212%29106%3A6%3C467%3ATAOC%3E2.0.CO%3B2-1

Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), Utility and Probability (pp. 15–18). Palgrave Macmillan. https://doi.org/10.1007/978-1-349-20568-4_5

Staal, M. A. (2004). Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework. https://human-factors.arc.nasa.gov/flightcognition/Publications/IH_054_Staal.pdf

Teehan, R., Clinciu, M., Serikov, O., Szczechla, E., Seelam, N., Mirkin, S., & Gokaslan, A. (2022). Emergent Structures and Training Dynamics in Large Language Models. Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, 146–159. https://doi.org/10.18653/v1/2022.bigscience-1.11

U.S. Nuclear Regulatory Commission (NRC). (1981). Checklist for Evaluating Emergency Procedures in Nuclear Power Plants (ML102560010). Washington, DC: U.S. NRC. https://www.nrc.gov/docs/ml1025/ML102560010.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30. arXiv: 1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Vinge, V. (1993). The Coming Technological Singularity: How to Survive in the Post-Human Era. Whole Earth Review, 81, 88–95. https://ntrs.nasa.gov/api/citations/19940022856/downloads/19940022856.pdf

Wolfram, S. (2002). A New Kind of Science. Wolfram Media. ISBN: 978-1579550080.