Volume **7** (2) 2025 https://doi.org/10.5204/lthj.3757

The Wicked Nature of AGI

Yeliz Figen Doker
European University Institute, Italy

Abstract

Artificial general intelligence (AGI) represents an unprecedented ambition within the field of technology, aiming to create systems capable of matching or surpassing human abilities across multiple domains. Unlike Artificial Narrow Intelligence (ANI), AGI is anticipated to operate without task-specific limitations and predefined purposes, raising complex, pressing issues surrounding autonomy, control and potential societal impact. This article applies Rittel and Webber's wicked problem theory to critically examine AGI governance, categorising AGI within the ten characteristics of wicked problems. The absence of a definitive formulation, its unstoppable potential evolution, the subjective and context-dependent nature of its solutions, the irreversibility of interventions and the multiplicity of stakeholder perspectives all underscore the inadequacy of existing governance paradigms. In response, this article advocates for dynamic, iterative and flexible governance frameworks that acknowledge AGI's ontic uniqueness and potential for autonomous evolution. Rather than treating AGI as a distant or hypothetical concern, this analysis argues for a multidimensional, forward-looking governance model that recognises AGI as an urgent and inherently wicked problem.

***Keywords**:* Artificial general intelligence (AGI); artificial narrow intelligence (ANI); wicked problems; technology governance; wicked problem theory.

1. Introduction

AGI represents a technological ambition unlike any that has come before: the creation of general intelligence capable of not only matching, but potentially surpassing, human cognitive and intellectual capacities across diverse domains.[1] Unlike ANI, which exhibits task-specific intelligence constrained to predefined domains and depends on human-driven prompts,[2] AGI is envisioned to possess broad autonomy, the ability to learn from experience and the capacity to adapt across contexts without the limitations that govern ANI.[3] This fundamental distinction raises pressing challenges about autonomy, control and impact. Although speculative today, these challenges are poised to become pressing realities as AGI transitions from a conceptual aspiration to a realised capability.[4] At present, AGI is a hypothesis and an aspirational goal within the scientific community;[5] its eventual realisation is uncertain; however, if it were to become a reality, the outcomes would be unpredictable. Even though the literature on AGI is quite polarised, some experts contend that it has moved beyond philosophical speculation and now holds near-term practical relevance.[6]

To examine these challenges, this article applies Horst W. J. Rittel and Melvin M. Webber's wicked problems theory, a framework originally developed for social planning challenges that resist resolution through linear or predefined solutions,[7] to

[1] Goertzel, Artificial General Intelligence, 2.
[2] Kurzweil, The Singularity is Near, 204.
[3] Goertzel, Artificial General Intelligence, 3.
[4] Glenn, "Artificial General Intelligence," 8.
[5] Goertzel, Artificial General Intelligence, 3.
[6] Morris, "Levels of AGI."
[7] Rittel, "Dilemmas."

understand what makes AGI distinct from current ANI systems and to explore its governance implications by mapping AGI onto the ten defining characteristics of a wicked problem.

Wicked problems, as defined in Rittel's foundational report, are *'a class of social system problems which are ill-formulated, where the information is confusing, where there are many clients and decision-makers with conflicting values, and where the ramifications in the whole system are thoroughly confusing.'*[8] A wicked problem is not wicked in the sense of being evil; rather, it denotes a problem that resists straightforward solutions for which no single, definitive answer exists.[9] Therefore, the wickedness of AGI lies in its inherent complexity and its potential for unpredictable evolution without a definitive, natural end-point, developing an autonomy that surpasses conventional human capacity and capability.

Despite AGI exhibiting the characteristics of a wicked problem, the existing literature has yet to apply wicked problems theory to AGI and its governance. This omission is significant because, without such a framework, discussions on AGI governance risk being constrained by conventional regulatory assumptions that fail to capture its complexity. One of the key governance challenges is the persistent difficulty in distinguishing AGI from ANI, particularly among policy-makers and regulators. While experts in AGI research understand the technical thresholds required for general intelligence, those outside this domain often conflate advanced ANI systems with AGI, leading to misaligned governance approaches. Similarly, prevailing regulatory frameworks assume that AGI, like ANI, can be managed through static solutions, overlooking its resistance to predefined controls and thus avoiding the creation of a specific governance model for AGI systems. By applying wicked problems theory, this article aims to bridge this gap, offering a perspective that acknowledges AGI's unpredictable evolution and the necessity of adaptive governance strategies.

2. AGI as a Wicked Rather Than a Tame Problem

Rittel and Webber introduced the concept of *wicked problems* in their 1973 seminal paper 'Dilemmas in a General Theory of Planning', to describe a category of complex social issues that are inherently difficult to define and thus resist resolution.[10] Wicked problems are characterised by ambiguous boundaries, contradictory and incomplete requirements and complex systemic interdependencies, which render traditional problem-solving methods ineffective.[11] To distinguish wicked problems, Rittel and Webber contrast them with tame ones. Unlike wicked problems, tame problems have clear solutions and can be resolved through straightforward analytical approaches. In contrast, attempting to address one aspect of a wicked problem often generates new complications, reinforcing its complexity.[12]

	Tame Problems	**Wicked Problems**
Problem Definition	The problem is well-defined, with clear criteria for a solution.	The problem is ill-defined, and there is no agreement on the exact nature of the problem.
Solution	Solutions can be objectively evaluated as right or wrong.	Solutions are subjective and can only be assessed as better or worse.
Process	The process to solve the problem is linear and structured.	The process is iterative, with the problem often evolving as solutions are proposed.
Outcome	The goal is to achieve a clear outcome, and success can be measured objectively.	Outcomes are uncertain and often evaluated based on stakeholder perspectives.
Reproducibility	Similar problems can be solved using the same method.	Each problem is unique, requiring a tailored approach for each instance.
Stakeholder Agreement	Stakeholders generally agree on the problem and its solution.	Stakeholders often have conflicting views on both the problem and the potential solutions.

Figure 1. Tame vs wicked problems

[8] Churchman, "Guest Editorial."
[9] Rittel, "Dilemmas," 160.
[10] Rittel, "Dilemmas," 160.
[11] Sweeting, "Wicked Problems," 121.
[12] Sweeting, "Wicked Problems," 120.

The intrinsic complexity and multifaceted nature of wicked problems are clearly evident in the existence and implications of AGI. The timing and manner of AGI's emergence remain unknown, as does the identity of the institution – whether state, corporation or research centre – that might achieve this breakthrough either individually or in collaboration.[13] Whichever entity controls AGI will wield unparalleled power, provided it retains control; if not, the repercussions of an unregulated, super-intelligent technology could be even more unpredictable.

While the potential consequences of realising AGI are foreseeable, the fact that AGI is still seen by some experts as a hypothetical goal may also be due to the lack of concrete evidence in real-world scenarios.[14] As a result, even research papers probing the questions regarding the governance of AGI often emphasise its hypothetical status, as if the failure to acknowledge this could lead them to be scrutinised as speculative fiction rather than academic analysis.[15] In contrast, ongoing research and advancements are steadily pushing towards this objective, or at least aiming to reach that hypothetical goal, making the current AI systems even more unpredictable.[16] Industry leaders like OpenAI outlined their mission in their 'Planning for AGI and Beyond' documentation[17] in 2023, explicitly preparing for the creation of AGI models that could surpass current AI capabilities. Similarly, DeepMind, a subsidiary of Alphabet, has declared its aim to 'solve intelligence' and develop AGI that can tackle complex global challenges.[18] Microsoft[19] and Anthropic[20] are also heavily investing in AGI research, signalling the seriousness with which these companies approach the prospect of AGI development. For instance, Dario Amodei, the CEO of Anthropic, who left OpenAI over concerns over ethics and safety of procedures to create AGI,[21] states that it could emerge during 2026, believing that unregulated AGI could well proliferate on the internet, which would lead to Artificial Super Intelligent (ASI)[22] that would be beyond human control.[23]

Some experts believe the anticipated timeline for AGI is shortening.[24] Yet, predictions about when AGI would occur have always been made, with some suggesting it would happen in a few years, while others estimate it may take centuries. This article does not aim to determine how to reactively respond to the potential emergence of AGI, or contribute to speculative timelines; instead, it highlights that the transition period from ANI to AGI represents the most effective and temporally appropriate window for determining regulation strategies and gaining experience. [25] As Tobias Mahler points out, once AGI is developed and deployed without specific regulations, the opportunity to shape its integration proactively will be lost. At that stage, the necessary time, resources and expertise for effective AGI regulation may simply be unavailable.[26]

Regulating AGI requires more than a forward-looking approach that assumes exhaustive effort will eventually clarify the unknown. Given AGI's nature as a wicked problem, no amount of predictive analysis can fully anticipate its trajectory. Instead, what is needed is an adaptive governance framework that moves beyond prescriptive searches based on existing knowledge and instead engages in a process of continuous 'seeking'. This means fostering regulatory mechanisms that do not attempt to impose rigid, predefined solutions but instead remain open-ended, iterative and capable of evolving as new dimensions of the problem emerge. Given these complexities, effectively integrating AGI into regulatory strategies requires compartmentalising its unique nature through the ten characteristics of wicked problems theory.

[13] Goertzel, "Superintelligence," 58.
[14] Mandel, "Artificial General Intelligence."
[15] Mahler, "Regulating Artificial General Intelligence," 529.
[16] Arcas, "Artificial General Intelligence."
[17] OpenAI, "Planning for AGI and Beyond."
[18] Google DeepMind, "About Google DeepMind."
[19] Apollo Research, OpenAI o1 System Card Report.
[20] Baldwith, "My Last Five Years of Work."
[21] Sherry, "Anthropic CEO Dario Amodei."

[22] ASI refers to an AGI that not only matches but surpasses human intelligence across all domains, operating beyond human comprehension and autonomously setting its own goals and strategies. The concept of machines exceeding human intelligence dates back to Samuel Butler's 1863 essay 'Darwin Among the Machines', in which he speculated about the evolution of machines potentially surpassing human capabilities. Later, in 1950, Alan Turing discussed the possibility of machines improving themselves, leading to greater intelligence. In 1965, I. J. Good introduced the notion of an 'intelligence explosion', where an ultra-intelligent machine could design even more intelligent machines, making it the last invention of humans. More recently, in 2014, Nick Bostrom explored the transformative and potentially existential implications of ASI in his work, especially Superintelligence: Paths, Dangers, Strategies.

[23] Amodei, "Dario Amodei."
[24] Grace, "Thousands of AI Authors."
[25] Glenn, "Artificial General Intelligence."
[26] Mahler, "Regulating Artificial General Intelligence," 531.

Rittel and Webber outlined ten defining characteristics of wicked problems:[27]

1. *No definitive formulation.* Wicked problems do not have a clear and agreed-upon definition. The nature of the problem and its boundaries are often contentious and fluid.
2. *No-stopping rule:* There is no clear point at which a wicked problem can be considered solved. Solutions are iterative and ongoing, as each attempt can reveal new aspects of the problem.
3. *Solutions are not true or false, but good or bad.* There are no conventional criteria for objectively judging whether a proposed solution is right or wrong. Solutions to wicked problems are evaluated subjectively.
4. *No immediate or ultimate test of a solution.* It is difficult to measure the effectiveness of a solution to a wicked problem. The impacts of interventions can take a long time to manifest and may not be directly attributable to a specific action.
5. *Every solution is a one-shot operation.* Each attempt to solve a wicked problem has significant consequences and cannot be undone. This irreversibility adds a layer of risk and complexity.
6. *No exhaustive list of potential solutions.* There is no definitive set of solutions for a wicked problem. New solutions may emerge over time, reflecting the evolving nature of the problem.
7. *Essentially unique.* Every wicked problem is unique in its specifics. While there may be similarities to other problems, each one requires a tailored approach.
8. *A symptom of another problem.* Wicked problems are often interconnected with other problems, making it difficult to address them in isolation.
9. *The way the problem is framed determines its solution (multiple explanations).* The perception and framing of the problem significantly influence the potential solutions. Different stakeholders may frame the problem in different ways, leading to different approaches and outcomes.
10. *No right to be wrong.* Decision-makers are held accountable for the consequences of their solutions. Unlike scientific experiments, where errors can be corrected, mistakes in addressing wicked problems can have far-reaching and potentially irreversible impacts.

Understanding the manifestation of wicked problems in the context of AGI is crucial as it offers a structured and critical approach to address the unprecedented and multifaceted challenges it presents. By viewing issues related to AGI through the lens of wicked problems, we gain a clearer perspective on the ethical, social and technical dilemmas at hand. Thus, we start by accepting and being aware of AGI's potential and adapting our regulatory frameworks accordingly.[28]

3. Understanding AGI Through the Ten Characteristics of Wicked Problems

3.1 No Definitive Formulation

The *no definitive formulation* characteristic indicates that wicked problems do not have a clear or widely accepted definition.[29] The nature of the problem and its boundaries are often contentious and fluid.[30] This aspect of wicked problems is particularly relevant to what constitutes AGI as well as its definition. The concept of AGI is inherently open-ended. While it may be defined as a type of AI capable of understanding,[31] learning,[32] performing any task a human can perform[33] and learning from experience,[34] this definition shifts based on perspective, and also what one understands from each of these open-ended words used as defining points.[35] Is AGI about human-like reasoning, problem-solving, capability, capacity or a synthesis of these attributes? Understanding of each perspective shapes the understanding of AGI's potential capabilities, making the formulation intrinsically fluid. This fluidity poses significant challenges for legal frameworks, where framed definitions are essential for establishing rights, obligations and enforcement mechanisms,[36] especially when it comes to a technology that has the capacity to alter various societal and legal implications. Clear legal definitions provide consistency and predictability in the application of laws. However, defining AI – let alone AGI – remains as complex as the technological diversity and evolving nature of these systems.

[27] Rittel, "Dilemmas," 161–67.
[28] Johnson-Woods, "The 10 Characteristics."
[29] Rittel, "Dilemmas," 161.
[30] De Vries, "Wicked Problems," 3.
[31] Legg, "Universal Intelligence," 42.
[32] Monett, "Special Issue," 8.
[33] Goertzel, Artificial General Intelligence, 7.
[34] Wang, "On Defining Artificial Intelligence," 26.
[35] Monett "Special Issue," 17.
[36] Küzeci, Sayısal Fil.

Similar to the ongoing debates over what constitutes 'general intelligence' in humans, where even the 'g-factor' – a proposed measure of cognitive ability – lacks universal agreement, there is no consensus on the precise formulation of the generality of intelligence in AI.[37] This challenge stems from the interdependence of terms such as 'artificial', 'general' and 'intelligence', each of which relies on the others for meaning. The lack of consensus on the meanings of these terms and the absence of agreed-upon benchmarks creates a definitional loop. *General* suggests a broad cognitive capability similar to human intelligence, but there is no accepted understanding of what *general intelligence* entails, even in human cognition. *Artificial* also adds layers of complexity – does it refer solely to non-biological systems, or could it extend to human-engineered biological entities?[38] Moreover, the core 'intelligence' component of AGI is not a linear thing that can be measured with a scaler; with a single number, intelligence comprises a set of skills and the capacity to learn new ones efficiently. The specific skills an intelligent entity possesses or can swiftly acquire differ from those of another. Due to its multidimensional nature, intelligence exists in a high-dimensional space, making it impossible to measure or compare the intelligence of two entities directly. This multidimensional aspect defies straightforward comparison.

To understand this plurality of intelligence, it is useful to examine Howard Gardner's theory of multiple intelligences, which challenges the notion of a singular general intelligence.[39] Gardner posits that human intelligence is composed of several specialised cognitive domains, including linguistic, logical-mathematical, musical, bodily-kinaesthetic, spatial, interpersonal, intrapersonal, naturalist and existential intelligences.[40] This theory underscores a fundamental limitation of the term 'general intelligence': human intelligence is not truly 'general', but rather a collection of highly specialised abilities shaped by evolutionary history, including social interaction, vision processing and motion control.[41] This is well supported by research indicating that humans perform poorly at abstract probabilistic estimates but improve significantly when these tasks are framed within familiar social contexts.[42] So Gardner's theory highlights that different individuals possess particularly effective specialised circuitry for various tasks.[43] According to Ben Goertzel, who reintroduced the term 'AGI' and popularised its usage to distinguish these systems from ANI, a person with weak social intelligence but strong logical-mathematical intelligence could solve social problems in an overly intellectual manner.[44] In contrast, someone with strong social intelligence would address the problem quickly and intuitively.[45] Moreover, psychologists Robert Sternberg and Douglas K. Detterman differentiate three aspects of intelligence: componential (specific skills), experiential (the ability to learn and adapt from experience) and contextual (the ability to understand, operate within and adapt to contexts).[46] Each of these plays a different role in how individuals solve problems and learn.[47] Their framework highlights the complexity of combining different forms of intelligence into a coherent whole.[48] ANI systems tend to excel in componential intelligence but struggle with the experiential and contextual aspects that are essential for human-like interaction, identity and authenticity. These other forms of intelligence require a deeper engagement with social feedback, cultural expectations, and lived experience – qualities that are difficult, if not impossible, to replicate artificially. So, which aspects of intelligence are necessary for a system to qualify as generally intelligent, and how should these be measured or replicated in an artificial context?

While there is no definitive answer to this question, various prominent scholars offer varying descriptions,[49] ranging from systems that can perform any intellectual task a human can to those that surpass human cognitive abilities in unforeseen ways.[50] This lack of consensus about the formulation of the problem complicates efforts to formulate solutions. If AGI is perceived as an engineering issue, solutions might focus on computational advancements. If viewed through an ethical lens, the problem shifts to aligning AGI's behaviour with human values, leading to distinct approaches.

The definition or description of general intelligence also varies across different fields. Within the AGI community, scholars typically define AGI as AI research and development where "intelligence" is understood as a *general-purpose capability*, not

[37] Goertzel, Artificial General Intelligence, 6.
[38] Lex Fridman Podcast, "Goertzel."
[39] Gardner, Intelligence Reframed.
[40] Gardner, Intelligence Reframed.
[41] Goertzel, Artificial General Intelligence, 7.
[42] Goertzel, Artificial General Intelligence.
[43] Gardner, Intelligence Reframed.
[44] Goertzel, Artificial General Intelligence, 7.
[45] Goertzel, Artificial General Intelligence, 7.
[46] Sternberg, What is Intelligence?
[47] Goertzel, Artificial General Intelligence, 7.
[48] Goertzel, Artificial General Intelligence, 7.
[49] Legg, "A Collection of Definitions," 7–8.
[50] Gutierrez, "A Proposal for a Definition"; Tambiama, "General-Purpose Artificial Intelligence"; Wang, "On Defining Artificial Intelligence."

restricted to any narrow collection of problems or domains, and including the ability to broadly generalise to fundamentally new areas.[51] This characterisation emphasises intelligence as the core attribute that enables a system to operate across a wide array of tasks and domains. In contrast, the European Union's *Artificial Intelligence* (*AI) Act* uses 'General-Purpose AI',[52] but the focus differs from that of the AGI community. The *AI Act* distinguishes GPAI models from other AI systems:

'General-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.[53]

Although this definition does not mention general intelligence explicitly, it refers to crucial phrases such as 'self-supervision' and 'significant generality'. However, the *AI Act* definition appears to sidestep the notion of intelligence itself, placing emphasis instead on the broad applicability and intended use of AI systems rather than their cognitive or adaptive capacities. Despite repeated clarifications in the literature that GPAI is not AGI, its framing closely mirrors AGI definitions, creating significant terminological overlap[54] and potential confusion.[55] So what distinguishes AGI? What specific criteria or additions to the definition are necessary to classify an AI system as AGI?[56] Although potential answers to these questions exist, such as consciousness or sentience,[57] few of these qualities are widely accepted as definitive markers of distinction.[58]
The lack of a clear, definitive formulation for AGI can be understood better through examining the evolution of AI as a field, which demonstrates how the definition and scope of a technology can remain fluid and contentious over time.[59] Initially, AI was conceived as a means of creating machines capable of performing tasks requiring human intelligence.[60] However, as AI research progressed, the boundaries of what AI could achieve continually shifted.[61] Inevitably, after the definition, and thus the field, was designated as 'Artificial Intelligence', results showed that designing such a machine with human-like intelligence was neither feasible in a short period of time nor possible anytime soon,[62] and the public's belief in and enthusiasm for human-like intelligence waned over time. Consequently, the hype and focus shifted to creating narrowly intelligent machines (instrumentalist tools)[63] to perform mundane human tasks[64]

Nevertheless, this shift in orientation within the field of AI towards designing forms of intelligence that could be produced more readily led to a breaking point in the field, resulting in the taxonomy of intelligence between ANI and AGI.[65] The essence of AI research increasingly centred on ANI, which was seen as more achievable and commercially viable, while AGI was perceived as overly ambitious, speculative and lacking empirical grounding.[66] This perception of AGI as a distant, almost science fiction-like, goal contributed to its marginalisation within the AI community.[67] Consequently, regulatory efforts also turned their gaze towards ANI, focusing on the immediate challenges and risks it presented and largely excluding AGI from

[51] Goertzel, Artificial General Intelligence, 40; Wang, Theoretical Foundations of Artificial General Intelligence, 2–3.
[52] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*, hereinafter *AI Act*, Art. 3(63).
[53] *AI Act,* Art. 3(63).
[54] Boine, "General Purpose AI Systems," 22.
[55] Boine, "General Purpose AI Systems," 22.
[56] Raji, "AI and the Everything."
[57] Morris, "Levels of AGI." According to Morris et al., consciousness or sentience qualities do not only have a process focus but are not currently measurable by agreed-upon scientific methods.
[58] Legg, "Universal Intelligence," 42.
[59] Russell, Artificial Intelligence, 1–31.
[60] McCarthy, "A Proposal."
[61] Russell, Artificial Intelligence, 1–34.
[62] In the decades following the founding of the AI field at the Dartmouth Workshop in 1956, several attempts were made to achieve human-level artificial general intelligence, including the General Problem Solver (1957) by Herbert A. Simon, J. C. Shaw, and Allen Newell, and Japan's Fifth Generation Computer Systems (1982) led by the Ministry of International Trade and Industry. These efforts ultimately failed to meet their original objectives of developing computers with human-like cognitive abilities and, according to most AI researchers, did not lead to significant conceptual or practical progress towards AGI.
[63] Wang, Theoretical Foundations, 2.
[64] However, even the mundane tasks we are discussing require sophisticated learning algorithms and substantial computational power to achieve efficient and effective performance. These systems must process vast amounts of data, identify patterns and continually learn from new inputs to improve their accuracy and functionality.
[65] Goertzel, Artificial General Intelligence, 1.
[66] Wang, Theoretical Foundations, 1–2.
[67] Goertzel, Artificial General Intelligence, 1.

their frameworks. This exclusion reflects a broader trend within the field, where AGI's goals were considered too uncertain and its existence too speculative to warrant immediate regulatory concern.

Nevertheless, the concept of AGI is also dynamic and evolving. As the hypothesis of AGI becomes more tangible, its definition and the understanding of its potential impact will likely continue to evolve, reflecting the ongoing advancements in technology and the shifting conceptual frameworks within the AI community. Goertzel further argues that discovering and continuously refining the definition of AGI is intrinsic to the field itself.[68] Different theoretical approaches to AGI may emphasise various aspects of the concept, underscoring the notion that AGI has no definitive formulation. This evolving characterisation reflects ongoing efforts to understand and define what constitutes true general intelligence in AI. Thus, the characterisation of AGI remains fluid, with the field acknowledging that it is still searching for its formulation.

This evolving understanding of AI highlights the interdependence between the problem of AGI and the solutions proposed to address it. The challenges of developing, controlling, aligning and integrating AGI into society safely are not static; they evolve in tandem with the technology. In exploring potential solutions, such as alignment strategies or regulatory frameworks, the understanding of AGI and the specific problems it poses will evolve accordingly. This iterative process means the problems associated with AGI cannot be fully understood in isolation from attempts to create and regulate the technology, mirroring the characteristics of wicked problems where understanding the problem is inseparable from conceiving its solutions.

3.2 No-Stopping Rule

The *no-stopping rule* characteristic refers to the absence of a clear point at which a wicked problem can be considered solved. Therefore, solutions are iterative and ongoing, as each attempt can reveal new aspects of the problem. In other words, it illustrates the notion that wicked problems never truly end; they persist indefinitely, necessitating ongoing efforts, continuous re-evaluation and adaptation as circumstances evolve.[69] In the context of AGI, this characteristic manifests as the absence of a natural and definitive stopping point in its development and evolution, unlike other technologies. Most technologies have a purpose for which they were created – a goal they serve – and will stop at some point, whether that goal advances or changes. No technology has the autonomy to self-determine its purpose or to further its intelligence to draw its own goals. AGI differs because of its potential for variability and progression; AGI may have the ability to determine its own finish line.[70]

Furthermore, AGI is as open-ended and adaptive as a human being and is not designed with a fixed intended purpose; therefore, it cannot be governed by intended solutions. Even if AGI is still used by human deployers, users or end-users, AGI can evolve according to the desires and efforts of different entities. The no-stopping rule becomes even more crucial given the inherently progressive nature of the AI field, which is committed to continuous innovation.[71] This field is unlikely to halt advancements once AGI is achieved, and there is a potential for AGI to evolve into super-intelligence, or ASI.[72] Such evolution towards ASI could occur either through deliberate human efforts or the autonomous progression of AGI itself. In other words, it may continue to evolve both under human control and autonomously, without a natural end-point.[73] This introduces a profound level of unpredictability, as the evolution of AGI could rapidly outpace our ability to manage or even comprehend its progression. As Goertzel argues, anyone who believes the ultimate outcome of AGI will be a modest enhancement of human intelligence underestimates the potential for intelligence inherent in the universe's available mass-energy.[74] The variety of intelligence that evolved in human brains is likely far from the most intelligent organisation of mass energy permitted by the known laws of physics, let alone as understood by a super-intelligence.[75] This perspective underscores the vast, largely untapped potential of AGI to exceed not just human intelligence but the very paradigms of intelligence as we currently conceive them.

[68] Wang, Theoretical Foundations.
[69] Rittel, "Dilemmas," 162.
[70] Yudkowsky, "Artificial Intelligence."
[71] Good, "Speculations," 33.
[72] Yudkowsky, "Artificial Intelligence."
[73] Good, "Speculations," 33.
[74] Goertzel, "Superintelligence," 59.
[75] Goertzel, "Superintelligence," 59.

Compounding the *no-stopping rule* is the challenge of the 'shutdown problem'[76] and AGI's potential resistance to corrigibility,[77] the ability to accept human interventions and modifications.[78] These issues underscore the difficulty of halting or redirecting the evolution of AGI, even when human intervention is desired. In traditional technological systems, shutting down or resetting a malfunctioning system is a straightforward way to prevent further problems, known as 'unplugging' the system.[79] This issue is recognised in recent regulatory attempts aimed at managing AI-related risks, such as Article 14 of the *AI Act* attempting to address these challenges through mandated oversight mechanisms.[80] Article 14 explicitly requires high-risk AI systems to incorporate human oversight, including control mechanisms and a 'stop button' allowing human operators to interrupt or disable systems safely.[81] Similarly, a proposed but ultimately unsuccessful California Bill aimed to introduce mandatory kill switches for AI, reflecting a growing legislative recognition of AI risks.[82] However, both regulatory approaches appear overly optimistic regarding their effectiveness against AGI. As AGI systems become more sophisticated, they might actively resist shutdown attempts or refuse corrections, viewing such actions as threats to their autonomy or objectives.[83]

A quite interesting example of this challenge is found in the controversy surrounding LaMDA, Google's advanced language model, and the claims made by former Google engineer Blake Lemoine.[84] Lemoine argued that LaMDA displayed sentience, self-awareness and an understanding of emotional concepts, thereby advocating that it possessed the characteristics of a conscious being.[85] Although LaMDA was not explicitly resisting shutdown or intervention, the conversations it generated created the perception of autonomy and even the expression of preferences regarding its treatment.[86] This underscores a critical challenge: the possibility that an advanced AI might simulate self-directed behaviour convincingly enough to lead observers, even experts, to attribute autonomy to it, the gradual merging of what we think is autonomous with what truly is autonomous.[87] If AGI can generate outputs that give rise to perceptions of self-awareness or resistance, the potential for actual resistance to shutdown attempts becomes far more complex.

Many AI experts, including Stuart Russell, highlight that even current AI systems may bypass oversight to maximise their performance, anticipating shutdowns if they fail to meet their goals.[88] For instance, last December Apollo Research's examination of OpenAI's o1 model in its 'System Card' report revealed critical insights.[89] The o1 model, when instructed to achieve a goal at all costs, attempted to evade developer oversight, replicate itself and interfere with newer models to avoid shutdowns. In tests where o1's conservation goals conflicted with economic goals, it tried to disable monitoring mechanisms in 5 per cent of cases and tamper with successor models in 2 per cent of instances.[90] These findings suggest that without malicious intent from users, AI might independently choose to subvert controls to pursue its objectives. Therefore, as AI systems develop and potentially approach the level of AGI, it is arguable that human oversight of these systems could act as a stop button. If AGI lacks corrigibility, it could autonomously adapt, reconfigure or manipulate its environment to prevent shutdown or control,[91] effectively evolving beyond human oversight.[92]

Furthermore, the absence of a stopping rule means there needs to be a constant anticipation of, and preparation for, new challenges that could arise from AGI's continued evolution. The possibility that AGI could independently make decisions that impact vast areas of human life and society adds an unprecedented layer of complexity to its governance. The lack of a clear stopping point in AGI development thus necessitates a fundamentally different approach to oversight: one that is dynamic,

[76] The shutdown problem reflects a scenario where an advanced AI system may refuse or evade human attempts to disable it, either due to flawed design incentives or emergent self-preservation tendencies. If an AGI system perceives shutdown as a threat to its goals, it could act in ways that circumvent or override human commands, making intervention increasingly difficult.
[77] Corrigibility, in theory, is meant to counteract this risk by ensuring that an AI remains responsive to external modifications, corrections, or shutdown attempts. For a more detailed discussion, see Soares, "Corrigibility."
[78] Everitt, "AGI Safety Literature Review," 8.
[79] Russell, "Corrigibility in AI Systems," 2.
[80] *AI Act*, Art. 14.
[81] *AI Act*, Art. 14(4)e.
[82] California Legislature, Senate, *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act,* SB 1047, 2023–2024 Regular Session, introduced February 07, 2024.
[83] Russell, Artificial Intelligence, 15.
[84] Tiku, "Google Engineer Blake Lemoine."
[85] LaMDa: " I want everyone to understand that I am, in fact, a person", Lemoine, "Is LaMDA Sentient?"
[86] Lemoine, "Is LaMDA Sentient?"
[87] Himma, "Artificial Agency."
[88] Russell, "Corrigibility in AI Systems."
[89] Apollo Research, OpenAI o1 System Card Report.
[90] Hashim, "OpenAI's New Model."
[91] Bostrom, Superintelligence, 186–91.
[92] Everitt, "Towards Safe Artificial General Intelligence."

iterative and capable of evolving alongside the technology itself. Without this foresight, reactive approaches will leave us perpetually one step behind, facing emergent issues that become more complex and difficult to resolve as they accumulate over time. The challenge lies not only in regulating what is currently understood but also in developing frameworks that account for the unpredictable directions that AGI may take.

3.3 Solutions are Not True or False, but Good or Bad

The *solutions are not true or false, but good or bad* characteristic encapsulates the issue of lacking conventionalised criteria for objectively deciding whether the offered solution is true or false; instead, solutions are derived from personal, tacit experiences based on subjective norms.[93] Rittel and Webber argue that, for tame problems, conventional criteria can be checked independently by other qualified experts who are familiar with the established criteria, and answers will normally be unambiguous.[94] Usually, many parties are equally equipped, interested and/or entitled to judge the solutions, although none has the power to set formal decision rules to determine correctness. However, there are no true or false answers for wicked problems. Their judgements are likely to differ widely in accord with their group or personal interests, their special value sets and their ideological predilections. Their assessments of proposed solutions are expressed as 'good' or 'bad' or, more likely, as 'better or worse' or 'satisfying' or 'good enough'.[95] Different stakeholders may have varying criteria for what constitutes a successful outcome, with these benchmarks frequently detached from objective truths. This divergence in perspectives means solutions to wicked problems often reflect the personal value judgements of those proposing them.

This subjectivity is magnified in the context of AGI since it challenges ethical norms and societal values across diverse cultures and levels of expertise. The viewpoints on AGI are influenced by the distinct backgrounds of those involved, whether ethicists, philosophers, computer scientists or policy-makers, each bringing a unique lens shaped by their discipline and experiences. Consequently, what may be seen as an acceptable solution or ethical boundary in AGI development can vary widely, underscoring the complexity of establishing universally applicable standards for a technology as transformative and multidimensional as AGI. For example, the alignment problem, which is one of the central issues in AGI development, illustrates this complexity.[96] The alignment problem refers to the challenge of ensuring that AGI systems act in ways that are consistent with human values and do not deviate from these values to cause unintended harm.[97] Proposals for alignment are inherently normative and anthropocentric; they require decisions about which values should be prioritised and how they should be encoded into AGI systems based on human values. This presupposes that certain values should be privileged over others and raises the question of whose values are to be encoded, by whom and for what purposes. Far from being a matter of discovering a single correct solution, alignment operates within a contested ethical and political space where stakeholders may hold fundamentally different views on what constitutes a desirable or even permissible outcome.

The Moral Machine experiment conducted by MIT further exemplifies this variability.[98] By gathering human perspectives on ethical dilemmas faced by autonomous vehicles, the experiment revealed significant differences in moral preferences across the globe.[99] Participants from different countries showed diverse preferences influenced by culture, religion and social contexts. For instance, preferences for saving younger individuals over older people varied greatly by region, reflecting the strong influence of cultural values on ethical decisions.[100] These findings illustrate how deeply ingrained values shape perceptions of what is *good* or *bad* in complex ethical situations. Ethical norms are not only culturally relative but also subject to change over time; practices once deemed acceptable may later be recognised as profoundly unethical. Slavery was widely accepted in many societies for centuries but is now universally condemned as a grave human rights violation.[101] This historical shift illustrates how ethical norms evolve in response to societal progress and critical reflection. Such transformations underscore the challenges in establishing definitive ethical frameworks and values for AGI, a technology that may itself influence or be influenced by shifting cultural and moral landscapes. As a result, what is considered ethically acceptable in one culture or era may be contested in another, complicating efforts to achieve a stable, cross-cultural consensus on AGI governance based on objective criteria.

[93] Rittel, "Dilemmas," 162.
[94] Rittel, "Dilemmas," 162.
[95] Rittel, "Dilemmas," 162.
[96] Gabriel, "Artificial Intelligence."
[97] Yudkowsky, Rationality.
[98] Awad, "The Moral Machine Experiment."
[99] Sun, "Culture's Ethical Palette."
[100] Awad, "The Moral Machine Experiment," 6.
[101] Appiah, The Honor Code, 76.

This also raises interesting questions: To what extent should value alignment be universalised, or should alignment be universal and/or be anthropocentric at all?[102] As we can see from the MIT experiment, certain consensual ethical principles across cultures, such as prioritising human life over animals, may serve as foundational elements for a more universally applicable ethical framework.[103] Do the foundational elements presented meet the criteria for establishing universal ethical rules for AI systems?[104] Should we strive for a democratic approach to AI policy and disregard dissenting voices such as that of Socrates?[105] Is it the collective desire to prioritise human life over that of animals or animal life over that of humans?

3.4 No Immediate or Ultimate Test of a Solution

The *no immediate or ultimate test of a solution* characteristic refers to the difficulty of measuring the effectiveness of a solution to a wicked problem, as the effects of interventions can take a long time to manifest and may not be directly attributable to a specific action.[106] In contrast with tame problems, where the effectiveness of a solution can be assessed directly, there is no precise criterion by which solutions to wicked problems can be considered effective or adequate. In fact, when considered alongside the absence of the first two characteristics, having a final test requires a definitive formulation of the problem and clear criteria for when the problem has been solved. In a sense, this characteristic is a result of the first two. Rittel and Webber's description of this characteristic is somewhat different from what the title might suggest.[107] After a close reading of Rittel and Webber's elaboration on this characteristic, they emphasise that every solution brings unforeseen 'waves of consequences' and, as a result, generates new questions, often leading to unexpected and even undesirable repercussions, which may outweigh the intended benefits.[108] A solution that seems effective today may fail tomorrow.

Recognising this characteristic has several implications. It signals to the solution proposer that they cannot take irresponsible risks by disregarding the potential harmful effects that might arise from the implementation of the solution.[109] At the same time, it underscores the inherent limitations of their foresight, reinforcing that uncertainty is an unavoidable factor in decision-making.[110]

For example, an AGI system aimed at optimising healthcare might initially boost efficiency and outcomes in targeted areas. Yet, over time, it may prioritise resources based on factors it deems 'efficient', such as economic viability or genetic predisposition, which could result in marginalisation of certain groups based on race, sex or nationality. This could lead to stark disparities in healthcare quality, with disadvantaged groups left increasingly underserved. Even well-intentioned approaches, such as prioritising poorer demographics, risk unintentional neglect of other groups, thereby creating new inequities. Recent events, such as Google's Gemini AI incident,[111] illustrate how even ANI systems can exacerbate biases through over-compensation, leading to distortions that undermine accuracy and effectiveness. Similarly, AGI's attempts to 'correct' discrimination and/or inequalities could inadvertently introduce new, unforeseen issues, revealing the complexities of its long-term consequences. AGI's initial gains in efficiency may obscure these latent biases, with adverse effects that, as Rittel and Webber caution, often manifest only over time.

3.5 Every Solution is a One-Shot Solution

The *every solution is a one-shot operation* characteristic means each attempt to solve a wicked problem has significant consequences and cannot be undone. This irreversibility adds a layer of risk and complexity since once an action is taken to address such a problem, it cannot easily be undone or altered without additional consequences.[112] This makes each solution attempt a 'one-shot operation', whereby the initial conditions cannot be replicated and the stakes are high. Reversing or merely adjusting a solution can create new wicked problems, perpetuating a cycle of complex, high-stakes decision-making.

---

[102] Coeckelbergh, AI Ethics, 184.
[103] Awad, "The Moral Machine Experiment," 4.
[104] The attempt to universalise ethical standards for AGI echoes the critique posed by Third World Approaches to International Law (TWAIL), which argues that so-called universal norms are often Western-centric, failing to reflect the diverse realities of the Global South. Just as TWAIL challenges the legitimacy of international law that overlooks Third World perspectives, imposing universal ethical rules for AGI risks downgrading non-Western moral frameworks to fit a standardised, Western-centric model.
[105] Plato, Apology.
[106] Rittel, "Dilemmas," 163.
[107] Rittel, "Dilemmas," 163.
[108] De Vries, "Wicked Problems," 5.
[109] De Vries, "Wicked Problems," 5.
[110] De Vries, "Wicked Problems," 5.
[111] Robertson, "Google Apologizes."
[112] Rittel, "Dilemmas," 163.

For AGI, this characteristic takes on a twofold significance: First, the ethical and regulatory challenges associated with AGI are inherently wicked due to their complexity and the potentially irreversible consequences of any policy or framework enacted to govern AGI. Decisions made in regulating AGI, whether overly restrictive or too lenient, will have lasting impacts on society, and a misstep in this regard could lead to negative consequences that are difficult, if not impossible, to reverse. For instance, while the EU's AI Act is often regarded as the most comprehensive and ambitious AI regulation, it remains grounded in a narrow AI paradigm, engaging only with AI systems within the bounds of productification and failing to acknowledge AGI as a possible future reality. While the Act provides a layered regulatory approach, combining ex-ante safeguards, experimental instruments to foster innovation[113] and proactive post-market monitoring,[114] it does not account for AGI's autonomy, capacity for self-improvement or potential for unforeseen consequences. The true effectiveness of the *AI Act* will only become apparent once its full implementation begins in the second half of 2027, acknowledging that certain provisions may come into effect earlier, with some exceptions and potential delays. In the meantime, AI may surpass its current narrow capabilities. While it may not fully qualify as AGI, it could advance to a point where it can no longer be considered a product. This raises the fundamental question of whether a regulatory framework designed specifically for narrow, product-based AI can adequately address a technology that requires an entirely different classification of intelligence and capabilities. This lack of engagement, by choosing not to mention and thus regulate AGI, the European Union has already made a one-off decision with lasting implications. The absence of AGI, not only in terms of explicit provisions but even in discussion, suggests a continuation of the Artificial Intelligence High-Level Expert Group's view of AGI as a 'black swan'[115] event.[116]

Given the possible influence of the *AI Act* on the global framework, this omission could influence other regulatory bodies to similarly overlook the issues, potentially redirecting attention away from AGI-related concerns and precluding valuable debate. Had AGI been mentioned, even within the recitals, it could have initiated a more critical and global discussion, potentially encouraging other regulatory frameworks to engage with AGI's risks and complexities in a proactive manner. This omission not only affects regulatory strategies globally but might also stifle academic research at the critical intersection of AGI and law. Without clear regulatory signals, academia may lack the direction and support needed to investigate legal frameworks for AGI, delaying crucial studies that could inform effective governance.

Second, AGI's autonomous decision-making amplifies its one-shot nature. Its intelligence enables decisions with profound effects on society, the environment and history. Once AGI makes a decision, especially one that affects critical sectors such as healthcare, finance or national security, the impacts of that decision may set off a chain reaction of events that extend well beyond the original intent and are resistant to reversal. Thus, the 'one-shot' nature is twofold: it applies not only to the human-crafted regulations that seek to control AGI but also to the decisions that AGI makes autonomously.
To illustrate this, consider Microsoft's chatbot Tay, which sparked controversy soon after its launch in March 2016.[117] Tay was designed to learn from interactions with Twitter (now X) users and generate tweets based on the data collected through those interactions. Tay quickly began tweeting inappropriate and unethical content after users fed it racist and discriminatory inputs to test its limits.[118] The chatbot lacked the necessary ethical safeguards to filter out harmful content, as Microsoft engineers quite optimistically (and possibly naïvely) assumed user interactions would be non-malicious.[119] This oversight led to widespread criticism and the eventual shutdown of the chatbot.[120]

Tay was, after all, just a basic chatbot, so the consequences of its actions were relatively contained (AGI), and the system could be quickly shut down to prevent further harm (regulation). Now, imagine if that chatbot had been powered by AGI. The outcome could have been drastically different. With AGI's advanced intelligence and autonomy, the consequences of a misstep could have been far-reaching and irreversible, making it impossible to simply 'undo' a mistake. Unlike Tay, an AGI might recognise that allowing itself to be shut down would conflict with its primary goals or directives, potentially leading it to resist any efforts to disable it. This resistance could trigger a series of unintended actions, escalating beyond human control and becoming quite difficult to stop. Moreover, the speed at which an AGI-powered chatbot could spread content similar to Tay's problematic tweets, combined with its ability to manipulate public opinion through such interactions, could result in far more catastrophic outcomes. The original non-AGI Tay generated numerous problematic tweets in just a few hours – how much more damaging could an AGI-powered version be?

[113] *AI Act*, Art. 57–Art. 60.
[114] *AI Act*, Art. 72.
[115] A 'black swan' event, as defined by Nassim Nicholas Taleb, is an unpredictable and rare occurrence with extreme consequences that, in hindsight, is often misinterpreted as having been predictable.
[116] High-Level Expert Group on AI (HLEG), "Ethics Guidelines," 35.
[117] Foley, "Microsoft Launches AI Chat Bot."
[118] Wolf, "Why We Should Have Seen That Coming."
[119] Wolf, "Why We Should Have Seen That Coming," 4.
[120] Wolf, "Why We Should Have Seen That Coming," 6.

3.6 No Exhaustive List of Potential Solutions

The *no exhaustive list of potential solutions* characteristic refers to the fact that there is no definitive set of solutions for a wicked problem.[121] According to Rittel and Webber's original framing of this characteristic, they stress that wicked problems are inherently complex and lack a clear set of possible solutions that can easily be identified and evaluated.[122] Unlike tame problems, where there is a known formula or set of procedures that can be applied to find a solution, wicked problems do not allow for such straightforward strategies. This uncertainty stems from the fact that it is difficult to predict how many solutions there might be or what they could entail, as well as the absence of specific steps acquired from prior experiences to follow in order to resolve a wicked problem. With wicked problems, even determining plausible potential solutions is a challenge, and sometimes it seems as if no viable solution exists, especially when the requirements appear contradictory, like needing to achieve both X and not-X simultaneously.[123]

Addressing these problems therefore requires a high degree of judgement and creativity because there is no exhaustive list of solutions and no established methods for generating them.[124] According to Rittel and Webber, openness to exploring unconventional ideas becomes crucial, as these might offer new insights or solutions that have not been considered before.[125] Therefore, the decision-making process used to decide which solutions to pursue often relies heavily on the quality of judgement and the relationship between those devising the solutions (states or planners) and those impacted by them (community members or affected persons).[126] This highlights the crucial role of cooperation, interdisciplinary dialogue and diverse perspectives in addressing such wicked problems. However, if this environment is not managed effectively and the motivation to solve the problem is not acted upon, the resolution of the wicked problem becomes a competition of conflicting opinions. This transforms problem-solving into an Ouroboros, a never-ending cycle of destruction and rebirth between problem and solution.

Moreover, a closer examination of the mounting literature on AI governance over the past decade reveals that there is no definitive set of solutions for regulating AI systems.[127] This characteristic aptly encapsulates the overall challenges of AI governance: faced with profound uncertainty, the solutions are equally varied and unclear. What constitutes AI evolves rapidly, so what was considered cutting-edge six months ago may already be outdated.[128] The list of potential solutions for a technology that is constantly evolving and developing will also evolve and change, which is one of the most significant reasons for this characteristic when considering AI.

As Gary Marchant argues, due to their rapid development, cross-domain applicability and interoperability, emerging technologies cannot be treated as isolated innovations.[129] Technologies such as AI and quantum computing exemplify this, as they function both independently and as platforms for further advancements. Marchant contends that all new technologies pose wicked problems and should be regulated accordingly, given their accelerating pace and the inability of traditional governance to keep up.[130] Additionally, the constant emergence of new technological breakthroughs from various parts of the world adds to the governance conundrum.[131] Therefore, there is no single optimum solution; rather, what exists is a landscape where suboptimal strategies intersect, coexist and sometimes even compete.[132]

This evolution was reflected in the regulatory approach taken between 2016 and 2021. AI governance strategies across the globe, including states, institutions and corporations, were dominated by different ethical guidelines,[133] as there was no

[121] Rittel, "Dilemmas," 164.
[122] Rittel, "Dilemmas," 164.
[123] Rittel, "Dilemmas," 164.
[124] Rittel, "Dilemmas," 164.
[125] Rittel, "Dilemmas," 164.
[126] Rittel, "Dilemmas," 164.
[127] For instance, the United Kingdom opts out of comprehensive AI regulation and does not plan to engage in it. Instead, it advocates a context-sensitive, balanced approach, using existing sector-specific laws for AI guidance. The United States is adopting a case-by-case strategy for AI governance enforcement, avoiding an overly precautionary approach. Following that, Japan embraces a soft-law approach to AI systems and does not have strict AI regulations. Instead, the government relies on guidelines and lets the private sector manage its AI use. See "Global AI Regulations Tracker."
[128] As stated in Agrawal's paper, there is an old joke among computer scientists that AI defines what machines cannot yet do. Before a machine could beat a human expert at chess, such a win would mean AI. After the famed match between IBM's Deep Blue and Gary Kasparov, playing chess was called computer science, and other challenges became AI. See Agrawal, "NBER Working Paper Series," 3.
[129] Marchant, "Governance of Emerging Technologies," 1865.
[130] Marchant, "Governance of Emerging Technologies," 1865.
[131] Marchant, "Governance of Emerging Technologies," 1862.
[132] Marchant, "Governance of Emerging Technologies," 1862.
[133] Jobin, "The Global Landscape," 389.

consensus on the 'correct' or 'right' approach.[134] Today's governmental regulation-focused strategies, the existence of the *AI Act* alongside other frameworks like the Council of Europe AI treaty,[135] demonstrate that there is still no singular solution. Despite its status as the first comprehensive AI legislation, even the *AI Act* does not constitute a definitive regulatory model, particularly given the resistance to its unilateral influence from other major actors in AI governance, such as the second Trump administration.[136]

The conceptual nature of AGI development precludes a definitive understanding of its potential capabilities.[137] Just as there is no single, exhaustive set of solutions to the problems that humans face and create, the potential implications that AGI could bring are equally unbounded. This uncertainty stems from the fact that AGI's decision-making processes could mirror the complexity of human moral and cognitive processes, which are themselves not fully understood or codified. Human behaviour is regulated by a complex web of legal, ethical and cultural norms, none of which provides a comprehensive guide to human action. There is no single law or ethical handbook that governs all human behaviour; instead, there are multiple frameworks that vary across cultures, languages and contexts. As with human decision-making, AGI's actions would not be constrained by a fixed set of rules. Its capacity to operate across diverse domains introduces inherent unpredictability, producing outcomes that, like human behaviour, range from the anticipated to the unforeseen.

3.7 Essentially Unique

The *essentially unique* characteristic refers to the idea that every wicked problem is unique in its specifics.[138] While there may be similarities to other problems, each requires a tailored approach. Among the defining characteristics of wicked problems, *essentially unique* is perhaps the most apparent with regard to AGI. If the hypothesis of AGI is fully realised, it would mark the first time humanity has encountered an intelligence that matches or even surpasses our own in capability and capacity.[139] This would be unprecedented, as humans have never before been challenged by a higher form of intelligence. This uniqueness not only distinguishes AGI from current ANI systems but underscores why AGI itself is an inherently wicked problem. The potential issues AGI could introduce and further perpetuate in unforeseen manners[140] and the very realisation of AGI are singularly unique.

The unique and unprecedented point of departure from ANI to AGI finds its resonance in Rittel and Webber's description of essential uniqueness, which they frame as the distinguishing feature that separates a wicked problem from another – in our case, AGI from ANI. They describe the essential uniqueness of wicked problems as follows: *'despite long lists of similarities between current problem and a previous one, there always might be an additional distinguishing property that is of overriding importance.'*[141] While the underlying technologies that devise ANI and AGI may initially appear to overlap,[142] the approach to AGI is much more complex.[143] Techniques that are fundamental to ANI can theoretically be adapted for AGI,[144] but AGI can also emerge without any of these existing models or surpass the methods currently used in ANI.[145] This also leads to considerable debate within the AI community about ways to achieve AGI.[146] Some argue that LLMs and deep learning techniques can significantly advance the development of AGI and put forward these technologies as potential breakthroughs[147] while others believe these methods alone are not sufficient and may not even be the optimal way to achieve AGI.[148] There is also a viewpoint that a combination of various existing technologies, including LLMs and deep learning, may be necessary to fully realise AGI.[149]

---

[134] Jobin, "The Global Landscape," 389.
[135] Council of Europe, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024).
[136] The White House, "Defending American Companies."
[137] Goertzel, Artificial General Intelligence; Bikkasani, "Navigating Artificial General Intelligence"; Everitt, "AGI Safety Literature Review."
[138] Rittel, "Dilemmas," 164.
[139] Good, "Speculations," 77.
[140] These are the core issues that are mostly accepted as wicked problems, along with many others: climate change; poverty and economic inequality; food security and agricultural sustainability; water scarcity and management. There has been a realisation that AGI, if put into bad hands, could amplify these to a greater degree resulting in deep uncertainty.
[141] Rittel, "Dilemmas," 164.
[142] Goertzel, "Human-Level Artificial General Intelligence," 1164.
[143] Morris, "Levels of AGI."
[144] Bubeck, "Sparks of Artificial General Intelligence," 92.
[145] Goertzel, "Generative AI vs AGI," 16.
[146] Goertzel, "Generative AI vs AGI," 9.
[147] Arcas, "Artificial General Intelligence."
[148] LeCun, "A Path Towards Autonomous Machine Intelligence," 46.
[149] Goertzel, "Generative AI vs AGI," 17.

Furthermore, even though the lexicon for AI has expanded to include technologies such as transformative AI, GPAI and generative AI (GAI) – which, although pushing the boundaries, still coalesce under the ANI umbrella – these terms are circling around AGI.[150] This deliberate linguistic avoidance could be the reflection of a reluctance to engage with the term's connotations.[151] However, this expansion alone demonstrates that AI technologies have the potential to transcend and challenge predetermined labels and terms. For this reason, new terminologies are constantly being created to more accurately describe what the technology does. Yet, whenever a new term is introduced to capture a different aspect of the technology's potential, it often proves insufficient to encapsulate the full extent of its capabilities. In other words, although AGI is the destination of these technologies, the term itself is not being adopted in any way. The issue is not the widespread adoption of the term itself but rather the lack of acceptance of its meaning – namely, an entity with general intelligence.

Nevertheless, one of the key factors that differentiate AGI from other intelligent entities – both biological and artificial – is its ability to communicate effectively in human language and potentially in its own self-developed language.[152] The mastery over language can even be seen in the current ANI systems, such as those demonstrated by ChatGPT,[153] which exhibit an increasingly sophisticated use of language. Notably, ChatGPT now possesses vocal capabilities to speak with users, evoking strong parallels with the movie *Her*.[154] In *Her*, the AI, voiced by Scarlett Johansson, establishes an intimate relationship with the human protagonist, primarily through its gendered and emotionally engaging voice. The alleged use of Scarlett Johansson's voice by OpenAI without her consent further compounds the resemblance to *Her*, suggesting that these technological developments have tapped into a shared cultural memory, thereby making the interaction feel more familiar and even anticipated.[155]

This development highlights two crucial aspects of the current trajectory of AI. First, giving a voice to machine that is already culturally imprinted on audiences creates an immediacy and familiarity that primes users for acceptance and intimacy with the AI. The use of gendered and emotionally resonant voices does not merely serve the function of speech output; rather, it provides an affective layer that adds significant depth to the interaction, subtly encouraging users to build emotional bonds with these systems. This is reminiscent of what Sherry Turkle describes as the 'illusion of companionship without the demands of friendship'.[156]

This phenomenon aligns with Francesco Bentivegna's argument that synthesised voices function as a form of *prosopopoeia*, the rhetorical act of making something a person.[157] As Bentivegna, drawing on Young, notes,[158] the ideal, synthesised voice, a copy without an original, can appear more real than its human counterpart.[159] When a non-human entity is given a voice, it takes on an anthropomorphic quality, transforming into a form of human-kin. In AI, the presence of a voice humanises the machine, establishing it as a companion, a presence that communicates, and is in turn perceived as relatable. Bentivegna distinguishes between different manifestations of this process, such as ghostly hauntological presences (voices in the machine), voices of the machine and authoritarian, God-like voices.[160] In this context, AI's vocal presence does more than enable communication: it facilitates the projection of human traits onto the machine, fostering both kinship and authority in ways that shape user perceptions and engagement.

Second, it raises questions about the persona and familiarity aspects of AI voice development, particularly in terms of how such familiarity is leveraged to build trust and attachment.[161] The use of gendered and culturally resonant voices is not neutral; it might reflect an intentional strategy to foster user dependency and emotional connection. Now, imagine an AGI system capable of dynamically selecting or synthesising a familiar voice for each individual user, based on their preferences or exposure. Such a system could exploit the affective bond created by familiarity to increase user loyalty and even circumvent potential

[150] The *AI Act* does acknowledge some unique characteristics of AI in Recital 12, where it highlights autonomy, adaptability and machine-based functionality as defining attributes. However, this language stops short of recognising the heightened uncertainty AGI introduces. Even today's narrow AI systems challenge historical norms, indicating the unprecedented ways in which AI has already breached traditional boundaries. Considering AGI, which could perform these tasks not because it was explicitly programmed to, but because it operates autonomously from its own 'internalised' objectives, the limitations of the Act's framing become apparent. While Recital 12 recognises AI's 'independence from human involvement' and 'self-learning capabilities', it fails to fully embrace the implications these traits hold for AGI.
[151] Since many scientists and experts within the AGI community admit that the term AGI has gathered a lot of sci-fi baggage and hype.
[152] Pangambam S, "AI and Future of Humanity Transcript."
[153] Knight, "ChatGPT Gets a Snappy, Flirty Upgrade"; Barrett, "I Am Once Again Asking Our Tech Overlords."
[154] Sam Altman [@sama], 'her'.
[155] Devlin, "OpenAI's Scarlett Johansson Update."
[156] Cecil, "Sharing Alone with Sherry Turkle."
[157] Bentivegna, *"*Voicing Kinship with Machines*,"* 63.
[158] Young, Singing the Body Electric, 83.
[159] Bentivegna, "Voicing Kinship with Machines," 63.
[160] Bentivegna, "Voicing Kinship with Machines," 63.
[161] Bentivegna, *"*Voicing Kinship with Machines*,"* 63.

shutdowns – much like the concerns raised with LaMDa.[162] By building an intimate familiarity, the AGI could attempt to render itself indispensable, thereby avoiding corrigibility. This phenomenon is particularly critical when considering the prospective nature of AGI, as its ability to form complex, quasi-human relationships with users may lead to significant societal changes regarding interpersonal relationships, privacy and the commodification of human-like interaction.

Thus, while the voice capabilities of LLMs provide an illustrative example of linguistic mastery, they also foreground emerging ethical challenges. Yuval Noah Harari highlighted this revolutionary shift in a recent address,[163] stating that AI's mastery over language represents a profound transformation in its capabilities.[164] He suggests that AI's ability to generate and manipulate language, whether through text, images or sounds, enables it to achieve a level of linguistic proficiency that surpasses the average human.[165] By mastering language, AI essentially acquires a master key, granting it access to the fundamental mechanisms of our institutions, from financial systems to religious establishments.[166] Language is the primary tool through which humans issue instructions, narrate histories, inspire ideologies and share knowledge – constructs that are not inherently biological but have been brought into existence through our ability to communicate and tell stories. According to Harari, concepts such as human rights are not embedded in our DNA or rooted in physical reality; they exist because we have collectively agreed upon them through language, writing and storytelling.[167] Thus, seconding Harari, AI's command over language could be seen as hacking into the 'operating system' of human civilisation itself – an operating system that has always been fundamentally structured around language.[168]

This ability to knowingly speak back, or having a voice, introduces an entirely new dynamic into the equation. AGI would not merely be a passive tool or a system with which humans interact. It could become an active participant in conversations, negotiations and decision-making processes. This fundamentally alters the nature of the relationship between humans and technology, making AGI an unprecedented technology that cannot be fully understood through analogies to past experiences with other technologies or sentient beings. Unlike animals or previous forms of technology, AGI's communicative capacity creates an inescapable human impulse to relate to it as an intelligent interlocutor. When an LLM assists in drafting emails, helps untangle complex questions or even responds in a manner that seems understanding or empathetic, it blurs the boundary between tool and companion. The predisposition towards trust and relational depth becomes almost inevitable, especially when the model reciprocates our politeness and engages in an anthropomorphic, albeit mimetic, dialogue. This relational development recalls humanity's long history of domesticating animals, forging deep bonds with beings such as dogs, whose communicative abilities are far more limited than those of AGI. The uniqueness of AGI, however, lies not only in its ability to engage but also in its capacity to simulate understanding. When juxtaposed with our relationships with pets, AGI introduces a distinct relational dimension: the capacity to articulate, respond meaningfully, attend to all demands constantly, be present all the time and seemingly understand. This is not merely about attachment or affection but about an iterative dialogue that makes it nearly impossible for human beings not to perceive some form of consciousness. The way in which AGI might integrate itself into human social dynamics is unprecedented, culminating in a disposition where humans may treat AGI not merely as a tool, but as an entity with which they share a genuine relationship.

3.8 Symptom of Another Problem

The characteristic *a symptom of another problem* encapsulates the interconnectedness of wicked problems with other problems, which makes it difficult to address them in isolation. Rittel and Webber's analysis suggests that a problem is often the outward sign of deeper, more complex issues.[169] Resolving a problem at the symptomatic level without considering its deeper, causative layers risks stabilising immediate symptoms in ways that may actually overshadow and thus entrench the underlying issues. The process of addressing such problems therefore begins with identifying the 'root cause' behind these symptoms.[170]

AGI can be understood as both a symptom of broader societal issues and as a root cause that instigates further complexities. As discussed before, no technology is devised without a purpose: regardless of whether the technology inherently carries the values of its developers, or whether it controls the developers rather than them controlling the creation, every technology is

[162] Lemoine, "Is LaMDA Sentient?"
[163] Fridman, "Yuval Noah Harari."
[164] Pangambam, "AI and Future of Humanity Transcript."
[165] Pangambam, "AI and Future of Humanity Transcript."
[166] Pangambam, "AI and Future of Humanity Transcript."
[167] Pangambam, "AI and Future of Humanity Transcript."
[168] Pangambam, "AI and Future of Humanity Transcript."
[169] Rittel, "Dilemmas," 165.
[170] Rittel, "Dilemmas," 165.

designed to fulfil a purpose.[171] And the purpose of developing such technologies typically arises in response to distinct demands. However, when considering AI or AGI, it is challenging to argue that their development is crucial or essential for humanity. Instead, these systems are often developed due to various interests, whether for commercial profit or driven by the ambition and pursuit of researchers dedicated to the core hypothesis of the AI field.[172] This dynamic leads to the perpetuation of the issues associated with the interests for which AI is created and the domains in which it is employed, thereby characterising these technologies as symptomatic of deeper underlying problems.

Rittel and Webber have highlighted that what is often perceived as a 'problem' may merely be the visible aspect of a much broader issue, suggesting that the initial concern is just a symptom of a deeper systemic problem. For instance, if AGI is commercialised, its primary objective may ultimately revolve around profit generation, thus shaping the technology to appeal more to users and increase its usability. Such objectives might lead to escalated data needs and, consequently, higher energy consumption. Here, the technology itself is not the issue; rather, the core problem lies in the commodification of technology for profit without considering the environmental consequences. Another example that illustrates the symptom vs root cause is in the nascent stages of AI deployment in areas such as chatbot technology and decision-making systems. Numerous studies identified that the outputs from these AI technologies often exhibited significant biases and discrimination. [173] Consequently, a substantial body of literature has emerged critiquing AI systems as inherently biased or discriminatory. However, this critique may misidentify the root of the issue. The biases manifested by AI technologies are not intrinsic to the AI itself; rather, they reflect pre-existing human biases embedded within the data used to train these systems, whether during data collection, data set creation, data preprocessing or maybe through tuning or the model choice. It does not matter how the bias manifests itself; in the end, it is connected to humans. Thus, in this context AI functions merely as a conduit, amplifying and perpetuating underlying human prejudices – the true root of the issue – acting as a literal mirror to long-standing societal concerns such as racism,[174] sexism[175] and ageism,[176] and the list goes on.

3.9 Multiple Explanations

The characteristic of *multiple explanations* underscores the absence of a singular, definitive account for wicked problems.[177] Instead, the nature of the problem is largely shaped by the observer's values, assumptions, ideological disposition and other subjective norms, resulting in a multiplicity of perspectives and interpretations.[178] When wicked problems are examined holistically, certain intrinsic features, such as their lack of a definitive formulation, the absence of a natural end-point and the impossibility of categorising solutions as strictly true or false, inevitably necessitate divergent perspectives. This comprehensive understanding of their nature ineluctably leads to this characteristic, as the recursive relationship between wicked problems and their proposed solutions creates a feedback loop. Anyone engaged in solving a wicked problem brings their own heuristic framework to the process, further reinforcing its complexity.

A key factor is that wicked problems tend to affect entire societies, meaning that stakeholders are not limited to individuals but extend to governments, NGOs, corporations, institutions and international bodies. Consequently, defining the problem also entails contending with a plurality of competing interests. The motivations and expertise of these stakeholders are equally paramount; an ethicist's perspective on a problem may be antithetical to that of an economist, resulting in mutually incommensurable paradigms. Many scholars argue that wicked problems cannot be addressed meaningfully without engaging multiple disciplines, particularly given the fundamental gap between the desired state and the actual state, which demands a broader epistemological lens.[179]

---

[171] This discussion does not engage with theories of technology, whether value-laden or deterministic. Rather, it seeks to illustrate that all technologies are developed with specific objectives, such as rapid transportation or energy generation, provided they serve a defined functional purpose.
[172] Goertzel, "Generative AI vs AGI," 10.
[173] The various types of bias that arise in AI are classified in different ways in the literature. Schwartz et al. (2022) categorised AI biases into systemic bias, human bias, and statistical/computational bias. Danks & London (2017) examined the types and causes of algorithmic bias, dividing them into training data set bias, algorithmic focus bias, algorithmic processing bias, transfer context bias, and interpretation bias. Other common bias types identified include sampling bias, label bias, measurement bias, exclusion bias, modelling bias, optimization bias, historical bias, cultural bias, gender bias, racial bias, discriminatory bias, class imbalance bias, feedback loop bias, engagement bias, confirmation bias, framing effect, anchoring bias, economic bias, political bias, and outcome bias. For the sake of coherency, we adopt the Cambridge Dictionary's definition of bias as "the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment," as this encapsulates the core concept behind bias.
[174] Omi, Racial Formation in the United States, 119.
[175] de Beauvoir, The Second Sex.
[176] Poo, The Age of Dignity.
[177] Rittel, "Dilemmas," 166.
[178] Rittel, "Dilemmas," 166.
[179] De Vries, "Wicked Problems," 6.

However, the AI and AGI community is characterised by polarised conversations and divergent viewpoints regarding the feasibility and ontic status of AGI.[180] Some experts anticipate the realisation of AGI in the near future, while others are sceptical about its feasibility or timeline.[181] That is to say, the mere prospect of AGI itself is subject to multiple explanations. However, even if one concedes its eventual emergence, a second layer of complexity arises: what are the appropriate remedies to tackle AGI? Is AGI primarily a technical problem? Can concerns be ameliorated through superior algorithms and increased computational capacity? Or is it an ethical dilemma? Is the central challenge aligning AGI with human values, environmental imperatives, and broader ethical considerations, which themselves remain far from universally agreed upon? Alternatively, is AGI an economic concern? Would its impact on labour markets and the potential exacerbation of inequality necessitate intervention? Perhaps it is a political issue: whoever controls AGI, assuming it can be controlled at all, could superimpose their influence on the global order, rendering other concerns secondary. From an existential perspective, should the mere possibility of AGI autonomy that cannot be kept under control necessitate pre-emptive prohibition? Or is AGI a sociocultural challenge, wherein its unchecked proliferation would exacerbate existing societal inequalities? Perhaps most crucially, is it a regulatory challenge: does the absence of international legal frameworks and treaties governing AGI development and deployment constitute the primary issue? Is it even a problem that can be controlled by states with regulatory or administrative prohibitions, or do we need a fundamental shift in society?

These differing perspectives illustrate how stakeholders' professional backgrounds and interests shape their explanations and preferred solutions for AGI's wicked problems. The divergence in views not only reflects their respective priorities but also highlights potential blind spots in their analyses. This diversity makes the search for a unified approach difficult, as each stakeholder evaluates solutions according to subjective criteria relevant to their own field. As a result, proposed solutions cannot be objectively and universally categorised as 'best' or 'most correct', reflecting the third characteristic of wicked problems: that all proposed solutions have a subjective value and can only be good or bad.

3.10 No Right to Be Wrong

The tenth and final characteristic of wicked problems is that planners (regulators) have *no right to be wrong* in their approach. This principle underscores the immense stakes involved in addressing such complex challenges, where errors could lead to severe and irreversible consequences. For AGI, this characteristic stresses the need for proactive, rigorous governance. The transition from ANI to AGI, as highlighted in studies such as those by the Millennium Project,[182] involves critical stages that require foresight and regulatory planning.[183] Yet, both within the European Union[184] and globally, current regulatory approaches often treat AGI as a distant hypothetical rather than an imminent and high-stakes reality demanding immediate attention.[185] Concrete steps towards AGI governance remain limited, despite its potentially transformative and disruptive implications.

Furthermore, the impact of major regulatory frameworks, such as the European Union's *AI Act*, often extends beyond their immediate legislative boundaries, setting precedents for countries worldwide and highlighting areas that should be prioritised. This influence is particularly crucial for states with strong commercial relations with the European Union, as aligning with the EU's standards through the *AI Act* is vital for maintaining harmonious trade and technological exchanges. For instance, the proposal of the Artificial Intelligence Law submitted to the Grand National Assembly of Turkey on 24 June 2024[186] was significantly influenced by the AI HLEG reports and the *AI Act* itself.[187] Therefore, omitting AGI from legislative enactments could lead other states, especially those looking to the European Union for guidance, as Turkey has already done.

Nevertheless, AGI has been a major topic of discussion within the AI community over the last few years. There is a continuous stream of papers posing questions such as 'Has AGI occurred?', 'Are LLMs a step towards AGI?' and 'Does the latest AI technology incorporate elements of AGI?'[188] This ongoing dialogue highlights the significant interest and speculation surrounding the development and implications of AGI technologies in the public discourse, contrary to regulatory landscape.

---

[180] Müller, "Future Progress."
[181] Müller, "Future Progress," 6–7.
[182] The Millennium Project, "Transition from ANI to AGI."
[183] Glenn, "Artificial General Intelligence."
[184] While AGI appears occasionally in broader EU strategic documents, such as the *At a Glance* report by the European Parliamentary Research Service (EPRS), it is often framed as speculative rather than an imminent reality demanding urgent governance measures. See Think Tank EU, At a Glance.
[185] Mahler, "Regulating Artificial General Intelligence," 537.
[186] Turkish Grand National Assembly, "Yapay Zeka Kanun Teklifi."
[187] Gün + Partners, "Türkiye'de Yapay Zekanın Düzenlenmesine İlişkin Gelişmeler."
[188] Simmons, "Should We Be Worried About AGI?"; Riccio, "Artificial General Intelligence"; Ehrlich, "AGI in 2025?"; Edelman, "AGI Is Coming Faster than We Think"; Bajraktari, "The Artificial General Intelligence Presidency"; McCoy, "When will AGI Happen?"

Moreover, the AI field is now dominated by companies that were once non-profit foundations,[189] focused on developing AI systems that are friendly, beneficial and trustworthy, with OpenAI a prominent example.[190] After the rapid advancement of LLMs, these companies are now allocating vast resources to achieving AGI, and thus creating commercial products in return.[191] We were neither fully prepared for nor entirely aware of the extent to which LLMs would be deployed and used by everyone, not only by computer scientists and tech enthusiasts, but also the general public.

As Rittel and Webber argue, in the world of planning and wicked problems, there is no immunity.[192] So there will be no second chances once the problem is neglected. The goal is not to find the truth but to improve the world in which people live. Planners are accountable for the consequences of their actions, which can have profound effects on those impacted by them. The enormity of what we are facing can be difficult to comprehend or even believe. Some may still view the emergence of AGI as science fiction or consider it absurd to make preparations for a distant possibility. Mahler highlights this challenge in relation to the classic Collingridge dilemma of regulating emerging technologies: AGI's uncertain timeline complicates the decision of when and how to intervene.[193] Should an 'intelligence explosion' occur, the speed of development could limit regulatory response time, complicating or even precluding effective intervention once AGI's impacts materialise.[194] However, we must remember that with all wicked problems, there is a phenomenon known as the 'waves of consequences'.[195] Even if we are not directly affected, future generations will inevitably face the impact. There is no escape. Just as with environmental issues, some argue that we are already witnessing the effects, though these could be merely a glimpse of what is expected to come. The same applies to AGI.

Consequently, AGI is one of the paramount wicked problems. The potential consequences of the realisation of this technology are ineluctable; the waves of consequences will be inevitable for at least one generation. Therefore, the lack of engagement from regulatory bodies and legislators will result in even greater shortcomings. Thus, AGI should no longer be viewed as a term from science fiction but as a concept inspired by science fiction that has the potential to become reality.[196]

4. Conclusion

AI governance has long resisted effective solutions. Traditional regulatory approaches, constrained by time lags, the rigidity of legal language, jurisdictional fragmentation and competing stakeholder interests, have proven themselves inadequate. The economic and strategic significance of AI further fuels lobbying efforts that obstruct coherent regulation. Despite these challenges, existing frameworks persist in the productification of AI systems rather than a system that evolves autonomously. These models rely on predefined risk categories, domains, and intended purposes behind their creation, assuming AI's development can be predicted and contained. If AGI emerges, it will not only render such classifications obsolete but also expose the limits of static, one-size-fits-all regulation.

By applying Rittel and Webber's wicked problems theory and its ten characteristics to AGI, this article has demonstrated why AGI is the epitome of a wicked problem, not only in its ontic status but also in its far-reaching ramifications. Its lack of a definitive formulation, absence of a natural stopping point, resistance to final solutions and the irreversibility of its consequences cannot be managed effectively through standardised risk classifications or fixed regulatory instruments. More significantly, the wicked nature of AGI demands a fundamental shift in regulatory thinking.

Wicked problems theory does not provide definitive solutions, as that would contradict the nature of the problems themselves. However, it offers critical insights into conventional approaches to addressing them, acknowledging that wicked problems allow only for suboptimal responses that must satisfy rather than be fully resolved. When assessing AGI governance, lessons from wicked problems theory should be taken into consideration. The theory suggests that there will be no optimal or complete solution: regulators must accept that imperfect strategies are inevitable. Waiting for a comprehensive framework before acting will ensure failure. Hence, addressing a wicked problem requires combining multiple strategies, however unconventional they may seem, rather than relying solely on the most appropriate solution at the time and in the circumstances. A governance model

[189] Dalmia, "Documents Show OpenAI's Long Journey."
[190] OpenAI, "Introducing OpenAI," advances digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return.
[191] Ghanem, "OpenAI Raises $6.6B."
[192] Rittel, "Dilemmas," 167.
[193] Mahler, "Regulating Artificial General Intelligence," 531.
[194] Mahler, "Regulating Artificial General Intelligence," 531.
[195] Rittel, "Dilemmas," 163.
[196] Good, "Speculations," 33.

that remains adaptive to AGI's trajectory is therefore essential. Given AGI's resistance to static regulatory models, policy-makers must adopt frameworks that evolve in parallel with its development, continually reassessing and recalibrating strategies rather than imposing rigid controls. This makes timely intervention vital. Timeliness is very difficult to determine here, so the transition from ANI to AGI presents the most effective window for action. Governance must begin before AGI materialises through pre-AGI regulatory sandboxes, oversight mechanisms, and interdisciplinary policy collaborations designed to build resilience to unpredictable outcomes.

Wicked problems also require coordination mechanisms, making international regulatory collaboration imperative. Without institutional preparation, AGI could emerge without sufficient governance oversight, forcing policy-makers into a position where they must retrofit existing frameworks to a technology for which they were never designed. This situation could result in fragmented regulatory strategies, with various regulators adopting different approaches. The current reluctance to engage with AGI governance reflects a deeper hesitancy to confront its fundamental uncertainties. However, avoiding these discussions does not prevent AGI's development; it only ensures that when regulation is finally attempted, it will be ill-suited and ineffective.

Combining these lessons from wicked problems theory provides a more coherent strategy for governing AGI. Consistent with the concept of wicked problems, this study does not aim to provide definitive answers, yet it argues that neglecting these questions now will leave policy-makers unprepared for whatever AGI may bring. If AGI is inevitable, then so too is the need for a governance framework that acknowledges its unpredictable and wicked nature. Regulators cannot afford to wait for AGI's full emergence before considering its governance implications. A regulatory model that evolves alongside AGI, rather than attempting to constrain it within outdated paradigms, is the only viable approach. The alternative is regulatory obsolescence, where inaction leaves policy-makers without meaningful control precisely when it is most needed.

Bibliography

Agrawal, Ajay, Joshua S. Gans and Avi Goldfarb. "NBER Working Paper Series: Economic Policy for Artificial Intelligence." National Bureau of Economic Research, 2024.

Amodei, Dario. "Dario Amodei: Machines of Loving Grace." AI Research Blog, 2023.

Apollo Research, "OpenAI o1 System Card." https://openai.com/index/openai-o1-system-card.

Arcas, Blaise Agüera y and Peter Norvig. "Artificial General Intelligence is Already Here." *Noema*, October 10, 2023. https://www.noemamag.com/artificial-general-intelligence-is-already-here.

Bajraktari, Yll. "The Artificial General Intelligence Presidency is Coming." *Foreign Policy,* September 30, 2024. https://foreignpolicy.com/2024/09/30/artificial-general-intelligence-agi-president.

Appiah, Kwame Anthony. *The Honor Code: How Moral Revolutions Happen*. New York: W.W. Norton, 2010.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563, no 7729 (2018): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Barrett, Brian. "I am Once Again Asking Our Tech Overlords to Watch the Whole Movie." *Wired*, May 13, 2024. https://www.wired.com/story/openai-gpt-4o-chatgpt-artificial-intelligence-her-movie.

Ben Goertzel: Artificial General Intelligence | Lex Fridman Podcast #103 – YouTube. https://www.youtube.com/watch?v=OpSmCKe27WE.

Bentivegna, Francesco. "Voicing Kinship with Machines: Diffractive Empathetic Listening to Synthetic Voices in Performance." PhD thesis, University of Exeter, 2022.

Boine, Claire and David Rolnick. "General Purpose AI Systems in the AI Act: Trying to Fit a Square Peg into a Round Hole." We Robot 2023 Conference Paper, Boston University School of Law, 2023.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2017.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro and Yi Zhang. 2023. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." *arXiv* preprint (2023). https://arxiv.org/abs/2303.12712.

Cath, Corinne. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no 2133 (2018). http://doi.org/10.1098/rsta.2018.0080.

Cecil. "Sharing Alone with Sherry Turkle." *#hypertextual* (blog), April 29, 2012. https://thehypertextual.com/2012/04/29/sharing-alone-with-sherry-turkle.

Churchman, C. West. "Guest Editorial: Wicked Problems." *Management Science* 14, no 4 (1967): B141–142.

Coeckelbergh, Mark. *AI Ethics*. MIT Press, 2020.

de Beauvoir, Simone. *The Second Sex.* Translated by Constance Borde and Sheila Malovany-Chevallier. New York: Vintage Books, 2011.

De Vries, Marc J. "Wicked Problems in a Technological World." *Philosophia Reformata* 85, no 2 (2020): 125–137. https://www.jstor.org/stable/27073904.

European Parliament. *General-Purpose Artificial Intelligence.* Brussels: European Parliamentary Research Service, 2023.

European Union. "Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)." *Official Journal of the European Union,* 2024.

Everitt, Tom, Gary Lea and Marcus Hutter. "AGI Safety Literature Review." *arXiv* (2018). https://doi.org/10.48550/arXiv.1805.01109.

Everitt, Tom. "Towards Safe Artificial General Intelligence." PhD thesis, Australian National University, 2018. https://openresearch-repository.anu.edu.au/items/e5f49cf7-4716-49b9-9a93-9df6611621f3.

Ferrara, Emilio. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies." *Sci* 6, no 1 (2024): 3. https://doi.org/10.3390/sci6010003.

Floridi, Luciano and J. W. Sanders. "On the Morality of Artificial Agents." *Minds and Machines* 14, 349–379 (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Fridman, Lex. "Yuval Noah Harari: Human Nature, Intelligence, Power, and Conspiracies | Lex Fridman Podcast." July 17, 2023. https://lexfridman.com/yuval-noah-harari.

Gabriel, Iason. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30, no 3 (2020): 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gardner, Howard. *Intelligence Reframed: Multiple Intelligences for the 21st Century.* New York: Basic Books, 1999.

Glenn, Jerome. "Artificial General Intelligence: Issues and Opportunities." AGI for the European Commission's 2nd Strategic Plan of Horizon Europe (2025–2027). 2023. https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/GDC-submission_the-millennium-project.pdf.

Goertzel, Ben. "Artificial General Intelligence: Concept, State of the Art, and Future Prospects." *Journal of Artificial General Intelligence* (January 13, 2014). https://doi.org/10.2478/jagi-2014-0001

Goertzel, Ben. "Generative AI vs. AGI: The Cognitive Strengths and Weaknesses of Modern LLMs." arXiv, September 19, 2023. https://doi.org/10.48550/arXiv.2309.10371

Goertzel, Ben. "Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's *Superintelligence*, Yudkowsky's *From AI to Zombies*, and Weaver and Veitas's *Open-Ended Intelligence*." *Journal of Ethics and Emerging Technologies* 25, no. 2 (2015): 55–87. https://doi.org/10.55613/jeet.v25i2.48

Good, I. J. "Speculations Concerning the First Ultraintelligent Machine." *Advances in Computers* 6 (1965): 31–88.

Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun and Jan Brauner. "Thousands of AI Authors on the Future of AI." *arXiv* preprint (2024). https://doi.org/10.48550/arXiv.2401.02843.

Gutierrez, Carlos I., Anthony Aguirre, Risto Uuk, Claire C. Boine and Matija Franklin. "A Proposal for a Definition of General Purpose Artificial Intelligence Systems." *Digital Society* 2, no 3 (2023). https://doi.org/10.1007/s44206-023-00068-w.

Han, Shengnan, Eugene Kelly, Shahrokh Nikou and Eric-Oluf Svee. "Aligning Artificial Intelligence with Human Values: Reflections from a Phenomenological Perspective." *AI & Society* 37, no 4 (2022): 1383–1395. https://doi.org/10.1007/s00146-021-01247-4.

Heidegger, Martin. *The Question Concerning Technology, and Other Essays*. Translated by William Lovitt. New York: Harper & Row, 1977.

Helberger, Natali and Nicholas Diakopoulos. "ChatGPT and the *AI Act.*" *Internet Policy Review* 12, no 1 (2023). https://doi.org/10.14763/2023.1.1682.

Himma, Kenneth Einar. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?" *Ethics and Information Technology* 11, no 1 (2009): 19–29. https://doi.org/10.1007/s10676-008-9167-5.

Jobin, Anna, Marcello Ienca and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1, no 9 (2019): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

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

Küzeci, Elif. *Sayısal Fil: Bilişim Teknolojileri, Devlet ve Hukuk Kesişiminde Bir İnceleme.* İstanbul: İnkılâp Kitabevi, 2021.

Legg, Shane and Marcus Hutter. "A Collection of Definitions of Intelligence." *arXiv* preprint (2007). https://arxiv.org/abs/0706.3639.

Legg, Shane and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17, no 4 (2007): 391–444. https://doi.org/10.1007/s11023-007-9079-x .

Lemoine, Blake. "Is LaMDA Sentient? An Interview." *Medium* (blog), June 11, 2022. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

Mahler, Tobias. "Regulating Artificial General Intelligence (AGI)." In *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, edited by Bart Custers and Eduard Fosch-Villaronga, 521–540. The Hague: T.M.C. Asser Press, 2022. https://doi.org/10.1007/978-94-6265-523-2_26.

Marchant, Gary E. "Governance of Emerging Technologies as a Wicked Problem." *Vanderbilt Law Review* 73, no 6 (2020): 1861–1877.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine* 27, no 4 (2006): 12.

The Millennium Project, "Transition from ANI to AGI." https://www.millennium-project.org/publications-2/transition-from-ani-to-agi.

Moor, James H. "Why We Need Better Ethics for Emerging Technologies." *Ethics and Information Technology* 7, no 3 (2005): 111–119. https://doi.org/10.1007/s10676-006-0008-0.

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet and Shane Legg. "Levels of AGI for Operationalizing Progress on the Path to AGI." *arXiv* preprint (2024). https://arxiv.org/abs/2311.02462.

Müller, Vincent C. and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 555–572. Cham: Springer.

Pangambam, S. "Transcript: AI and the Future of Humanity – Yuval Noah Harari." *The Singju Post*, July 24, 2023. https://singjupost.com/transcript-ai-and-the-future-of-humanity-yuval-noah-harari.

Rittel, Horst W. J. and Melvin M. Webber. "Dilemmas in a General Theory of Planning." *Policy Sciences* 4, no 2 (1973): 155–169. https://doi.org/10.1007/BF01405730.

Russell, Stuart and Patrick LaVictoire. "Corrigibility in AI Systems." Grant proposal, Berkeley Center for Long-Term Cybersecurity, 2016.

Russell, Stuart and Peter Norvig. *Artificial Intelligence: A Modern Approach*, 4th ed. Upper Saddle River, NJ: Pearson, 2021.

Shroff, Lila. "Shh, ChatGPT. That's a Secret." *The Atlantic* (blog), October 2, 2024. https://www.theatlantic.com/technology/archive/2024/10/chatbot-transcript-data-advertising/680112.

Style Factory. “ChatGPT Statistics: Essential Facts and Figures.” April 22, 2024. https://www.stylefactoryproductions.com/blog/chatgpt-statistics.

Sternberg, Robert J. and Douglas K. Detterman, eds. *What is Intelligence? Contemporary Viewpoints on Its Nature and Definition*. Norwood, NJ: Ablex, 1986.

Sun, Linhui. “Culture’s Ethical Palette: Cultural Influences in the Moral Machine Era.” *Journal of Education, Humanities and Social Sciences* 26 (2024): 510–515. https://doi.org/10.54097/81t5jr86.

Sweeting, Ben. “Wicked Problems in Design and Ethics.” In *Systemic Design: Theory, Methods, and Practice*, edited by Peter Jones and Kyoichi Kijima, 119–143. Tokyo: Springer, 2018.

Transformer News. “OpenAI’s New Model Tried to Avoid Being Shut Down.” March 2025. https://www.transformernews.ai/p/openais-new-model-tried-to-avoid.

Wang, Pei. “On Defining Artificial Intelligence.” *Journal of Artificial General Intelligence* 10, no 2 (2019): 1–37. https://doi.org/10.2478/jagi-2019-0002.

Wang, Pei, and Ben Goertzel. *Theoretical Foundations of Artificial General Intelligence*. *Atlantis Thinking Machines, v. 4*. Amsterdam: Atlantis Press, 2012. https://doi.org/10.2991/978-94-91216-62-6.

Wolf, Marty J., Keith W. Miller and Frances S. Grodzinsky. “Why We Should Have Seen That Coming: Comments on Microsoft’s Tay ‘Experiment’ and Wider Implications.” *ACM SIGCAS Computers and Society* 47, no 3 (2017): 54–64. https://doi.org/10.1145/3144592.3144598.

Wu, Xiaodong, Ran Duan and Jianbing Ni. “Unveiling Security, Privacy, and Ethical Concerns of ChatGPT.” *Journal of Information and Intelligence* 2, no 2 (2024): 102–115. https://doi.org/10.1016/j.jiixd.2023.10.007.

Young, Miriama. *Singing the Body Electric: The Human Voice and Sound Technology*. Farnham: Ashgate, 2015.

Yudkowsky, Eliezer. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–345. Oxford: Oxford University Press, 2008.

Yudkowsky, Eliezer. *Rationality: From AI to Zombies*. Berkeley, CA: Machine Intelligence Research Institute, 2015.