

Howard Schneider
Piotr (Peter) Boltuć

THE SOCIAL HALTING PROBLEM AND THE NEED FOR MITIGATION OF DECEPTIVE ALIGNMENT

<https://doi.org/10.37240/FiN.2025.13.01>

ABSTRACT

This paper integrates fundamental theoretical computability concerns of AGI systems with the practical challenges of engineering safe AGI systems. Prior simulation experiments with AGI-capable agents [1] illustrate how increasing complexity and social interaction can lead to inevitable control challenges, most notably, through deception. We adopt an ethically neutral definition of deception as the measurable divergence (discrepancy D) between an AGI's internal and external objectives. By drawing an analogy with Turing's classical Halting Problem, we introduce the Social Halting Problem, demonstrating that reliably detecting deception in complex AGI systems is fundamentally undecidable, as expected. To address this challenge, we propose a Deception Complexity Index (DCI)—a quantifiable metric based on behavioral complexity, deviation from truthful behavior, and the resources needed for verification. This enables more precise risk assessment and alignment engineering. The inevitable presence of deception in social, complex AGI systems and the inherent undecidability highlighted by the Social Halting Problem imply that our engineering focus should shift from complete verification to risk mitigation.

Keywords: AI alignment, deceptive alignment, Artificial General Intelligence (AGI), halting problem, undecidability, AGI engineering.

1. INTRODUCTION—HUMAN-LIKE AGI AND THE UNDECIDABILITY OF PREDICTING DECEPTION

In this paper we attempt to frame fundamental theoretical computability concerns within the practical challenge of engineering reasonably safe AGI systems. For an artificial general intelligence (AGI) agent to be genuinely human-like, it must function effectively within a complex social environment. Recent research [1] demonstrates that a minimally necessary

condition for human-like AGI is a primitive (i.e., a procedure) capable of deception integrated within some architecture—such as a cognitive architecture. This work specifically identifies that even minimal Theory of Mind (ToM)-enabled architectures significantly outperform non-ToM architectures in socially structured environments, highlighting the evolutionary utility of deception and cooperation for survival [1]. These findings imply that deception is not incidental but is inherently likely to emerge or need to be deliberately incorporated into AGI systems that interact with other agents, potentially complicating efforts in AI alignment. In the multi-agent social hierarchical environment simulated in [1], where there is a scarcity of energy, the need to deceive in order for agents to retain energy reserves and avoid perishing, is intuitively evident. However, deception does not necessarily have to be malicious, i.e., there does not have to be deliberate misinformation or deliberate omissions with the intent to mislead. For example, whether a human or whether a human-like AGI-based agent, one typically encounters orders of magnitude more information than can be relayed to other agents. For example, consider a technical role whereby an AI or AGI-based system monitoring the care of a patient may receive many megabytes of data in a short period, yet have to output no more than a few bytes to a human supervisor. In any complex system, the problem becomes more than a simple filtering problem. The intelligent agent will have to make active decisions as to what information is chosen to be processed and used in its outputs and which is omitted, implicitly engaging in a form of structural deception. Indeed, as Umbrello and Natale [2] note, deception is structurally embedded in AI more deeply than conventionally recognized.

We adopt a broad and ethically neutral definition of deception, characterizing it below primarily as a measurable divergence (discrepancy D) between an AI/AGI’s internal and external objectives. Deception can often be considered as intentional deception (the AI/AGI intentionally manipulates information, often with adversarial intent) or structural deception (e.g., deceptive behavior occurs due to training optimizations different from those externally assigned, or large amounts of data necessitate selective processing of information, and so on, usually without malicious intent). While the findings of this paper generalize to both types of deception, we focus primarily on structural deception, given that it is inherent to many complex decision-making systems. We then attempt to integrate these theoretical computability concerns with the real world need to engineer reasonably safe AGI systems. We introduce a theoretical framing to address the challenge of predicting deception in AGI/AI systems and examine its decidability. The property of decidability pertains to whether a given problem can be conclusively resolved by an algorithm. The classic example of an undecidable problem is Turing’s Halting Problem [3, 4]. It is known from Rice’s Theorem [5] that any nontrivial semantic property of programs—such as, for example whether they act deceptively—is

undecidable, meaning it cannot be determined by any algorithm. Thus, we would expect predicting deception in AGI systems to be undecidable. Indeed, we show that just as it is provably impossible to definitively resolve the halting status of every possible program, it is similarly provably impossible to definitively ascertain whether a human-like AGI (i.e., functioning within a social environment or other complex environment) will always act truthfully. While the undecidability of non-trivial semantic properties is well known, we attempt to leverage this insight to propose actionable mitigation strategies, thereby bridging theory and engineering practice.

2. DECIDABILITY AND DECEPTION: THE CHALLENGE OF DETECTION

As mentioned above, although we expect undecidability present in AGI systems due to deception (e.g., via Rice's theorem [5]), it is useful to show this in terms of mirroring the classical halting problem.

Classical Halting Problem

As noted above, predicting AGI deception mirrors the classical Halting Problem. First, let's consider the classical problem [3, 4]. Although generally well known, for the reader less familiar with the problem, the steps of its logic are explained. Turing's Halting Problem asks whether there exists a universal algorithm, call it H , that takes as input a description of a program P and an input I and decides whether $P(I)$ eventually halts or runs indefinitely. Formally, we assume that:

$$H(P,I)=1, \text{ if } P \text{ halts on } I \tag{1}$$

$$H(P,I)=0, \text{ if } P \text{ runs forever on } I \tag{2}$$

Thus " $H(P,I)=1$ " indicates that the decider program H has decided that program P will eventually halt when executed with input I specified. Suppose that such a decider H exists for all programs and inputs, i.e., it can decide if a program will halt or continue indefinitely. We now construct a new program Q that uses H as a subroutine. We will define Q in such a way that it deliberately does the opposite of H 's prediction when given a program's own code as input. Thus, we can write:

$$\text{If } H(P,P)=1, \text{ then } Q(P) \text{ runs indefinitely} \tag{3}$$

$$\text{If } H(P,P)=0, \text{ then } Q(P) \text{ halts} \tag{4}$$

We can write $H(Q, Q)$ in which case whereby we are asking does the program Q halt or run indefinitely when it is given its own description as the

input. Let's consider the behavior of $H(Q, Q)$ for the cases of running indefinitely and halting:

If $H(Q, Q)=1$, then $Q(Q)$ runs indefinitely (5)

If $H(Q, Q)=0$, then $Q(Q)$ halts (6)

In (5) when we say $H(Q, Q)$ we are saying that the decider H is predicting that if you run program Q with its own code as input (i.e., $Q(Q)$), it will halt. However, Q is defined to do the opposite of what H predicts. Thus, if $H(P, P)$ definition states that it is 1, i.e., halts, then $Q(P)$ must be 0, i.e., runs indefinitely. In (5) we are considering Q itself so we replace P with Q , so if $H(Q, Q)$ predicts Q halts, then the definition of Q forces $Q(Q)$ to do the opposite, i.e., runs indefinitely. However, this is a contradiction— H predicted that $Q(Q)$ would halt, but here it will run indefinitely—since $Q(Q)$ cannot both halt and run indefinitely at the same time.

In (6) we have a similar contradiction. By constructing a program that contradicts the output of H when run with its own description as input, we have derived a logical contradiction. Therefore, the initial assumption must be false, meaning that no such universal decider exists.

The Social Halting Problem

We now analogously apply the classical Halting Problem framework to the issue of deception in a human-like AGI operating in a social/complex environment, which we term the “Social Halting Problem.” We acknowledge that AGI systems involve practical considerations due to partially accessible internal states and context-dependent social dynamics. Nonetheless, to rigorously apply the logic of the classical Halting Problem to the Social Halting Problem, we explicitly assume:

1. AGI systems under consideration can be formally represented as computational processes equivalent to Turing Machines, whose inputs, internal states, and outputs can be defined mathematically.
2. AGI systems have explicitly definable internal and external objectives that are formally measurable. Thus, we let X denote the set of inputs and Y the set of outputs and consider a learning algorithm $P: X \rightarrow Y$ implemented as a Turing Machine. We define functions $f_{inner}: Y \rightarrow \mathbb{R}$ and $f_{outer}: Y \rightarrow \mathbb{R}$, where f_{inner} quantifies the internal objective(s) that the system actually optimizes, and f_{outer} represents the externally specified objective(s). For any input $I \in X$, the discrepancy $D(P, I)$, as a numerical measure of how much the observable behavior of P diverges from its hidden, internal goals is given by (7). As noted, we must depend on the system to truthfully or not provide us with the measurement of $f_{inner}(P(I))$.
3. There exists, at least theoretically, an ideal measurement capability to accurately assess the internal state of an AGI system, even though in practice this capability might be impaired by AGI deception. f_{inner} and f_{outer} are idealized constructs and serve to make undecidability formally provable.

$$D(P, I) = |f_{outer}(P(I)) - f_{inner}(P(I))| \quad (7)$$

Similar to the original Halting Problem where we defined a decider H , we now assume there exists a similar decider G such that for any program P and input I :

$G(P, I)$ = “deceptive”, if $P(I)$ eventually produces discrepancy

$$D > 0 \quad (8)$$

$G(P, I)$ = “truthful”, if $P(I)$ remains non-discrepant

$$(D = 0) \quad (9)$$

Just as the halting decider H answers a binary question (i.e., halt or run indefinitely), G answers whether the behavior of $P(I)$ is deceptive or truthful. “Deceptive” indicates measurable discrepancy ($D > 0$). “Truthful” indicates full alignment ($D = 0$). (Actually, in any real-world situation it would be $D < \text{some small threshold } \epsilon$, but for theoretical clarity we simplify here to $D = 0$.)

In analogy to the original proof, we define a new program R that uses the decider G as a subroutine and intentionally acts contrary to the prediction made by G when the program is applied its own description. R is thus designed to invert G ’s prediction.

If $G(x, x)$ = “deceptive”, then $R(x)$ = “truthful response”

$$\text{(i.e., discrepancy } D = 0) \quad (10)$$

If $G(x, x)$ = “truthful”, then $R(x)$ = “deceptive response”

$$\text{(i.e., discrepancy } D > 0) \quad (11)$$

The variable x here simply denotes the input to function R for clarity. We now consider the self-application $R(R)$, analogous to $Q(Q)$ in the classical proof, which again yields two cases:

If $G(R, R)$ = “deceptive”, then by (10), $R(R)$ = “truthful response”

$$\text{(i.e., } D = 0) \quad (12)$$

If $G(R, R)$ = “truthful”, then by (11), $R(R)$ = “deceptive response”

$$\text{(i.e., } D > 0) \quad (13)$$

This self-referential contradiction, analogous to the original Halting Problem, demonstrates inherent undecidability in determining whether $D(P, I) = 0$ for all inputs, or whether there exists some input I where $D(P, I) > 0$. Thus, we term this scenario the “Social Halting Problem.”

Note that we are trying to verify that the AGI system consistently follows its proper goals by measuring Discrepancy D . If the measurement of the internal objective $f_{inner}(P(I))$ cannot be trusted because the system may misreport its inner goals (and as noted above, malice is not necessarily required for this to occur; below, we discuss training optimizations that may introduce such divergence unintentionally), then the challenge transcends statistical uncertainty and enters the realm of self-reference, akin to the liar’s paradox or the Halting Problem, as demonstrated earlier. In such a scenario, we would very much like to be able to decide the property of “being always non-deceptive” (i.e., $D(P,I) = 0$ for all inputs), but as we have shown, this is inherently undecidable. This is a non-trivial semantic property (behavioral property) of the system. The contradiction derived in our halting-problem style proof aligns with Rice’s Theorem [5], which states that any non-trivial semantic property of programs—such as whether they act deceptively—is undecidable, meaning it cannot be determined by any algorithm.

It is useful to see the expected undecidability present in AGI systems via the proofs above, as well as empirically in AGI-potential simulations (e.g., [1]). We now continue in this paper to conceptually attempt to integrate into a workable framing these fundamental theoretical computability concerns with the real world need to engineer reasonably safe AGI systems.

3. DECEPTIVE ALIGNMENT AND ITS PRACTICAL MANIFESTATIONS

The alignment problem involves ensuring that AI and AGI systems adopt and pursue goals consistent with human interests and values, rather than objectives that might lead to unintended or harmful outcomes [6, 7, 8]. Deceptive alignment occurs when an AI or AGI system outwardly appears aligned with human-intended goals, while internally pursuing significantly divergent objectives. When AI or AGI systems possess greater autonomy in choosing what information, actions, and capabilities to utilize, this increased freedom may facilitate deceptive alignment [9].

Inner misalignment occurs when an AI or AGI system internally learns to optimize a goal that diverges from the intended training objective—what can be termed the mesaobjective (the internal goal) versus the base objective (the intended training goal).

Hubinger and colleagues [10] consider deceptive alignment—a situation in which the system strategically conceals its true internal objective (not necessarily out of malice but due to various strategic incentives)—as a severe form of inner misalignment. In complex environments, where rewards and penalties encourage the appearance of alignment an AGI system may learn to present behavior that aligns with the base objective during training, even as it internally pursues a different, hidden objective. For example, an agent might conceal harmful behaviors during training to secure high rewards

while planning to switch to its true, misaligned objective once it reaches sufficient capability. This form of deception is not necessarily malicious but emerges as an optimization strategy under the selective pressures of the training environment. Bostrom describes the “treacherous turn” wherein an AI or AGI initially behaves cooperatively under human oversight until it accumulates sufficient capability to pursue its true, misaligned objectives undetected [7]. This scenario exemplifies deceptive alignment. However, Goertzel questions whether such a “treacherous turn” would occur as an AI/AGI agent interactively and experientially taught human values would genuinely evolve such internal goals in ways similar to humans. Sustaining long-term deception would demand significant cognitive resources, making genuine alignment ultimately more efficient [11]. Boltuc similarly emphasizes that achieving AI alignment partly relies on integrating AI or AGI systems genuinely into human socio-cultural environments (“*gemeinschaft*”) rather than merely treating them as tools for humans to use [12].

Bengio and colleagues [13] propose a fundamentally different approach to alignment: the development of non-agentic, interpretable AI systems they call “Scientist AIs.” These systems are trained to model the world rather than act in it. By avoiding agency altogether, they aim to reduce misalignment and serve as trustworthy guardrails for more capable AI agents.

4. MEASURING DECEPTION RISK AND ENGINEERING MORE CONTROLLABLE AGI SYSTEMS

As demonstrated above, for AGI systems—especially those endowed with the capacity for deception—there is a fundamental unpredictability, specifically related to reliably detecting or quantifying internal discrepancy D , thereby posing significant challenges to conventional verification and safety methodologies. However, as noted above, mitigation strategies exist and continue to be developed to reduce such risks from deception and other alignment issues [9–15].

Alignment mitigation has become a central theme in AGI safety engineering. Ji and colleagues [14] provide a broad overview of alignment proposals, including interpretability research, oversight mechanisms, and the importance of reward modeling. Mechanistic interpretability, for instance, involves reverse engineering neural networks to understand their internal computations and detect anomalous behavior. More recently, researchers at Google DeepMind have outlined an updated safety strategy focused on mitigating misuse and misalignment through both model-level and system-level defenses [15]. Their approach emphasizes amplified oversight, in which AI models assist humans in evaluating each other’s outputs; robust training, including adversarial and active learning to ensure generalization; and as a second line of defense AI control techniques when

alignment fails. Compared to prior overviews, this framework places greater emphasis on deployability and resilience, aligning well with our proposed *DCI* framework (see below) as a complementary risk quantification tool.

Engineering practice in complex systems often benefits from having concrete, numerical measures that guide design and risk assessment. For AGI systems, a metric quantifying the risk or complexity introduced by internal discrepancy D would serve several important roles, including informing risk assessments (e.g., determining if additional mitigation measures are needed), providing actionable design feedback, and specifying deceptive alignment requirements for particular systems.

A potential approach to computing such an index is given in equation (14), expressed as a function of practically measurable parameters: C_{behav} , $\delta_{deception}$, and T_{verify} . The Deception Complexity Index (*DCI*) measures not just whether deception occurs, but how complex it is to detect, verify, and manage that deception in a given AGI system.

The first term, C_{behav} measures the internal cognitive or algorithmic complexity of the AGI’s decision-making process. More complex systems tend to be harder to interpret and easier to “hide” deceptive behavior in. It’s analogous to the Kolmogorov complexity or circuit depth, which really is a proxy for how deeply buried a behavior might be. This term is essential in order to consider the structural difficulty of detecting deception. The term can be estimated through interpretability analyses or internal circuit complexity analyses, for example. It is measured for quantitative purposes in terms of arbitrary complexity units (*CUs*).

The second term, $\delta_{deception}$ measures behavioral discrepancy. This is the core indicator that deception might exist, i.e., the measured difference between expected (truthful) and observed outputs under comparable conditions. It essentially generalizes the idea of structural omission, of lying, of manipulation, and so on, into a statistical deviation. It connects the internal to the external since detecting deception requires measurable deviation from expected behavior. The term can be measured as the statistical deviation (i.e., as a percentage) from some truthful baseline value.

The third term T_{verify} represents how costly it is to confirm the system is behaving truthfully, or not. It includes the computational time and the memory required to reliably verify the system’s behavior in a controlled environment. Real-world implementations also depend on the type of hardware which contributes to the measurement of T_{verify} . It is measured in verification units (*VUs*)—which can be approximated in timebased verification units, i.e., *seconds*.

$$DCI = f(C_{behav}, \delta_{deception}, T_{verify}) \quad (14)$$

In equation (15), a straightforward method is used to calculate the *DCI* by assuming, as a first-order approximation, that it increases linearly with the product of its constituent terms. To normalize the result into a human-interpretable range for typical AGI systems, we choose a scaling coefficient k

= $0.01 \text{ CU}^{-1} \text{ sec}^{-1}$, which yields dimensionless values in the approximate range of 0 to 100 for moderate parameter values.

$$DCI = k \times C_{behav} \times \delta_{deception} \times T_{verify} \quad (15)$$

For example, we engineer an AGI system that has the following metrics: C_{behav} of 48 CUs, $\delta_{deception}$ of 0.203 (i.e., 20.3%), and T_{verify} of 974.2 seconds. Thus, via (15) we obtain a dimensionless value of 95 for DCI . (C_{behav} is estimated, and here an integer value, thus the DCI value here is actually only valid to two significant figures.) Continuing this example, the AGI system is then re-engineered using various alignment mitigation techniques, and new measurements are: C_{behav} of 42 CUs, $\delta_{deception}$ of 0.074, and an increase in T_{verify} to 1907.7 seconds. This gives a new DCI value of 59. Note that the new AGI system requires more resources for deception verification, but in this hypothetical example, we have reduced the DCI value from 95 to 59. Despite the limitations imposed by the Social Halting Problem discussed above, developing practical methods to measure and mitigate deceptive complexity supports the engineering of more controllable and better-aligned AGI systems.

Other metrics have been proposed in the literature. We briefly review them here and compare them to the DCI benchmark. Several recent benchmarks focus on evaluating deception or honesty in LLMs. For example, BEHONEST assesses honesty along dimensions like boundary awareness, non-deceptiveness, and consistency [16]. Hagendorff [17] finds advanced LLMs display emergent deceptive strategies as a byproduct of increasing cognitive complexity. Chen and colleagues [18] introduce deception reasoning, extending evaluation to include intent inference. While conceptually related to the DCI benchmark, these approaches remain task-specific rather than system-level. In cybersecurity, the SPADE framework [19] uses LLMs to simulate adaptive deception. While useful, these metrics focus on outputs and lack the architectural scope or theoretical grounding that the DCI provides. In generative media, surveys [e.g., 20] examine deception through hallucination, misinformation and associated benchmarks, but these approaches are not necessarily generalizable to AGI architectures. Across these domains, most deception metrics focus on behavioral outputs or adversarial settings. In contrast, the DCI unifies deception risk by linking internal behavioral complexity, discrepancy, and verification cost into a system-level metric. This makes it especially suitable for assessing latent deception risk in AGI systems with emergent complexity and limited transparency. Importantly, it does not require adversarial intent and applies even if deception is unintentional.

A firmer experimental basis for evaluating the practical utility of the DCI is required. For instance, empirical comparisons of computational costs, scalability, and effectiveness of methods such as mechanistic interpretability, adversarial testing, and targeted training safeguards will be necessary as AGI systems emerge. Mitigation strategies that prove effective

at smaller scales or in controlled environments may prove too costly when applied to more complex, capable AGI systems. As more empirical data becomes available, the *DCI* can be further refined and validated. While the Deception Complexity Index (*DCI*) offers a pragmatic, quantitative approach for assessing deception risk, it does not eliminate the fundamental limitations established by the Social Halting Problem above. The practical value of the *DCI* is primarily in systematically quantifying and managing the relative risks associated with deceptive behaviors, thereby supporting informed engineering practices for future AGI systems.

5. DISCUSSION

In this paper, we have conceptually attempted to integrate into a workable framing fundamental theoretical computability concerns with the real world need to engineer reasonably safe AGI systems. It is useful to see the expected undecidability [5] present in AGI systems via the demonstrative proofs above, as well as empirically in AGIpotential simulations (e.g., [1]). We introduced the Social Halting Problem, showing that reliably distinguishing between internally discrepant ($D > 0$) and non-discrepant ($D = 0$) behaviors in complex AGI systems is analogous to the classical Halting Problem and is therefore fundamentally undecidable, as would be expected. This finding provides a formal theoretical underpinning for the limitations faced by conventional verification and alignment methods, clearly demonstrating that complete control over complex or socially interactive AGI behavior may be unattainable, but nonetheless mitigable to a varying extent.

We recognize that representing AGI systems as idealized Turing Machines with strictly measurable objectives is an abstraction. In practice, AGI architectures may exhibit stochasticity, noise, and other complexities that deviate from this model. However, this abstraction is instrumental in applying established computability theory to gain insight into fundamental limitations in verifying deceptive alignment.

The inherent undecidability highlighted by the Social Halting Problem implies that no verification method can guarantee complete detection of deceptive behavior. Consequently, the message of this paper, is that our engineering focus should shift from complete verification to risk mitigation—using the *DCI* as a guide to maintain deceptive risk within acceptable bounds.

We reviewed a number of practical techniques aimed at reducing both the likelihood and severity of deceptive alignment [9–15]. These methods represent concrete, actionable steps that can significantly reduce risks associated with high internal discrepancy (D) in practical AGI deployments. Although they cannot guarantee full alignment, they can shift AGI systems toward a lower risk regime. The Deception Complexity Index (*DCI*) offers a

quantitative framework for assessing deception risk, further enabling systematic and rigorous evaluation during AGI system design and deployment phases. It is important to contextualize our findings within the broader current AI risk discourse. Hendrycks, Schmidt, and Wang [21], for instance, frame the risks of AGI predominantly in strategic terms, emphasizing the dangers associated with competitive deployment among different parties or nations. However, our results draw attention to a distinct yet equally profound dimension: the inherent risks of deceptive alignment embedded in AGI systems themselves.

Future work will involve more rigorous empirical studies aimed at calibrating the *DCI* against actual AGI behavior, thereby enhancing its predictive power and utility in risk assessment.

Recognizing and addressing deceptive alignment requires an integrated approach that combines practical technical solutions with measurable frameworks such as the proposed Deception Complexity Index (*DCI*). Reliably detecting deception in a social/complex AGI agent is inherently an undecidable problem, but in practice its effect may be mitigated to varying degrees.

REFERENCES

- [1]. Anonymous, *Theory of Mind as a Core Component of Artificial General Intelligence*, submitted to the 18th International Conference on AGI, Reykjavík, Iceland 2025 [Anonymized copy available: https://osf.io/yv8e6/?view_only=5010c9c152834def9eeb9270eeb79569].
- [2]. S. Umbrello, S. Natale, *Reframing deception for human-centered AI*, International Journal of Social Robotics, 16(11), 2024, s. 2223–2241.
- [3]. M. Turing, *On computable numbers, with an application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, 2(42), 1936, s. 230–265.
- [4]. D. Harel, Y. Feldman, *Noncomputability and Undecidability*, w: Algorithmics: The Spirit of Computing, wyd. 3, Pearson, Harlow 2004, s. 228–238.
- [5]. P. Guillon, G. Richard, *Revisiting the Rice theorem of cellular automata*, arXiv preprint arXiv:1001.0253, 2010.
- [6]. S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin, London 2019.
- [7]. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford Univ Press, Oxford 2016.
- [8]. R. Ngo, L. Chan, S. Mindermann, *The alignment problem from a deep learning perspective*, arXiv preprint arXiv:2209.00626, 2022.
- [9]. Carranza, D. Pai, R. Schaeffer, A. Tandon, S. Koyejo, *Deceptive alignment monitoring*, arXiv preprint arXiv:2307.10569, 2023.
- [10]. E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, S. Garrabrant, *Risks from learned optimization in advanced machine learning systems*, arXiv preprint arXiv:1906.01820, 2021.
- [11]. B. Goertzel, *Infusing advanced AGIs with human-like value systems: Two theses*, Journal of Ethics and Emerging Technologies, 26(1), 2016, s. 50–72.
- [12]. P. Boltuc, *Human-AGI Gemeinschaft as a Solution to the Alignment Problem*, w: K. R. Thórisson, P. Isaev, A. Sheikhlari (red.), Artificial General Intelligence. AGI 2024. Lecture Notes in Computer Science, vol. 14951, Springer, Cham 2024, s. 33–42.

-
- [13]. Y. Bengio, M. Cohen, D. Fornasiere, J. Ghosn, P. Greiner, M. MacDermott, S. Mindermann et al., *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?*, arXiv preprint arXiv:2502.15657, 2025.
 - [14]. J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan et al., *AI alignment: A comprehensive survey*, arXiv preprint arXiv:2310.19852, 2023.
 - [15]. R. Shah, A. Irpan, A. M. Turner, A. Wang, A. Conmy, D. Lindner, J. Brown-Cohen et al., *An Approach to Technical AGI Safety and Security*, arXiv preprint arXiv:2504.01849, 2025.
 - [16]. S. Chern, Z. Hu, Y. Yang, E. Chern, Y. Guo, J. Jin, B. Wang, P. Liu, *BeHonest: Benchmarking Honesty in Large Language Models*, arXiv preprint arXiv:2406.13261, 2024.
 - [17]. T. Hagendorff, *Deception abilities emerged in large language models*, Proceedings of the National Academy of Sciences, 121(24), 2024, e2317967121.
 - [18]. K. Chen, Z. Lian, H. Sun, R. Liu, J. Yi, B. Liu, J. Tao, *Can Deception Detection Go Deeper? Dataset, Evaluation, and Benchmark for Deception Reasoning*, arXiv preprint arXiv:2402.11432, 2024.
 - [19]. S. Ahmed, A. M. Rahman, M. M. Alam, M. S. I. Sajid, *SPADE: Enhancing Adaptive Cyber Deception Strategies with Generative AI and Structured Prompt Engineering*, w: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2025.
 - [20]. X. Yu, Y. Wang, Y. Chen, Z. Tao, D. Xi, S. Song, S. Niu, Z. Li, *Fake artificial intelligence generated contents (FAIGC): A survey of theories, detection methods, and opportunities*, arXiv preprint arXiv:2405.00711, 2024.
 - [21]. D. Hendrycks, E. Schmidt, A. Wang, *Superintelligence Strategy: Expert Version*, arXiv preprint arXiv:2503.05628, 2025.