

AGI: The AI Alignment Problem

Tejas Kamtam

1 Introduction

The AI Alignment problem is perhaps the most critical issue of the current century. With the rate of AI advancement year-on-year, Artificial General Intelligence (AGI)¹ is right around the corner and will likely pose an existential threat to humanity if misaligned. The alignment problem is the issue of certifiably ensuring AI/AGI understand and innately behave with human values in mind when making decisions toward their objectives (Strickland, 2023). However, this problem is far more insidious than it may appear. Questions of “How do we define human values?,” “How do we know we’re not being deceived?,” and “Is the AI just cooperating temporarily to achieve some nefarious task in the future?” are just a sample of the many genuine concerns that must be answered before the advent of transformative AGI.

1.1 Inner and Outer Alignment

To this end, many scientists and researchers have already begun to answer some of these questions - usually resulting in more questions. Paul Christiano (OpenAI), Jan Leike (Anthropic), and Eliezer Yudkowsky (MIRI) are a few of the top names in the AI safety landscape, each ensuring their respective organizations aid the development of safe and robust AI. Their research has divided the alignment problem into two major topics: inner and outer alignment (often referred to by their contrapositive, “misalignment”). Outer alignment relates to agents making decisions/actions that are perceivably aligned with human values. Inner alignment, on the other hand, considers whether the intentions and thought process (usually termed “chain-of-thought”) of the agent that came to its decision/action are truly aligned (Christiano, 2018). The contrapositive is generally easier to interpret: inner misalignment is present in a model “if [the] implicitly represented reward function doesn’t match the desired reward function” (Leike, 2022) of a sufficiently intelligent AI.

¹ AGI - “a machine capable of behaving intelligently over many domains” (Davis et al., 2009)

These two topics have resulted in a handful of AI safety subfields to tackle the overarching problem. Precisely, capabilities² and control evaluations— currently conducted by many startups, including MIRI³, Redwood Research, and FAR AI (personal communication, 2024) —to determine whether an AI is outer misaligned and mechanistic interpretability research to lift the veil over the black box shrouding our understanding of LLMs to determine inner alignment (a few larger organizations pursuing this problem include Google DeepMind and Anthropic; personal communication, 2024). However, there is still much to be discussed: timelines, the possibilities of superintelligence, challenges in the field outside of research, setbacks, etc.

1.2 Timelines

An essential question for understanding the threat level is how far away AGI is. Significant evidence shows that the alignment problem will be a primary concern for the next 5-100 years.

In her paper, commonly referenced as “Bio Anchors,” Ajeya Cotra attempts to project timelines of transformative AI⁴ using biological anchors as a framework for model complexity (Cotra, 2020). Cotra draws a comparison between model parameters and neurons in the brain. This analogy is supported by the shared improvement in intelligence with increasing brain size (number of neurons) and model size (number of parameters).

With this comparison, the current SOTA⁵ decoder models (autoregressive chat completion models like ChatGPT) show equivalent intelligence to that of a small rat. With this anchor in place, Cotra defines transformative AI as a model with as many parameters as the human brain has neurons. Extrapolating computational power, researcher counts, and model architecture growth, Cotra projects the appearance of

² AI research is generally classified as capabilities (advancing model performance and ability) or safety research

³ MIRI - Machine Intelligence Research Institute

⁴ Transformative AI, as Cotra defines, is an AI with sufficient intelligence to cause a transformation in technological advancement equivalent in impact to the industrial revolution

⁵ SOTA - State of the art

transformative AI in the next 50 years with a probability of ~48% and in the next 75 years with a probability of ~70% (Karnofsky, 2021). Although these timelines are indeed quite alarming, it is essential to note that these timelines make many assumptions about the continued growth of computational power and model architecture. Most importantly, however, is the anchor assumption as new open-source models and novel architectures like MAMBA (Gu and Dao, 2023) are counterexamples to the intelligence and parameter count correlation.

2 Concerns

Beyond the issue of malevolent actors, a perilous problem with the example of AlphaFold⁶ in the hands of terrorists, there are many problems associated with misalignment. Consider an artificial human-level intelligence; what would such an agent be capable of? Shallow issues arise with machiavellian, narcissistic, malevolent behavior and actions. Historically, just a single evil person can cause mass destruction, pain, and suffering to millions; imagine an agent capable of accessing the internet, cracking the best encryption algorithms, impersonating powerful world leaders, and generating deep fakes and defamatory statements. However, some more pressing concerns AI Safety researchers are focused on relate to the fundamental nature of AI development and AGI behavior — many of which are incredibly simple yet impossibly tricky to solve.

2.1 The Shutdown Problem

A natural, naive approach to preventing devastating consequences from misaligned AI would be just to press the “shutdown button.” However, this proves to be much more complicated than it would appear. Consider that even a simple AI-powered Roomba would learn to prevent shutdown to complete its floor cleaning objective. Inherently, AIs are rewarded for completing their objectives, which would necessitate preventing shutdown from an external source (Ngo, 2022).

A misaligned AI would share this motivation to prevent a shutdown; however, it may optimize to do the most “bad” possible before getting shut down. A sufficiently intelligent AI may understand that it will be shut down should it exhibit misaligned behavior and (if misaligned) instead try to hide its malevolent actions from the user while maximizing its misaligned objective. An even more intelligent AI may even be able to deceive its users or convince them that shutting down will prevent a negative outcome by “playing dead” or interacting intentionally insidiously.

⁶ AlphaFold is a SOTA⁴ protein generation and classifier developed by Google DeepMind. The model has helped make groundbreaking discoveries in protein chemical structure. In the wrong hands, AlphaFold could create gene-targeting viruses, virulent and deadly diseases, etc.

2.2 Ontology Identification

As training an AI to comprehend human language and basic logical notions is possible, researchers propose teaching a model of human values may be possible to ensure its alignment. This is precisely the ontology identification problem: establishing the foundational concepts an AI should enforce to engage with the world. This involves delineating how concepts, relationships, and properties are represented and considered during decision-making (Anon., 2015). However, this endeavor has its challenges.

Firstly, subjectivity permeates the process. Ontologies often hinge on a human perspective on entities and their interrelations, a view that varies across individuals and cultural contexts. Defining a solid, discrete, and distinct set of core “human values” is an arduous task — a problem grounded in moral philosophy more than ML (Taylor, 2023). Secondly, enforcing actions aligned with the model's instilled ontology is convoluted to evaluate consistently, as decisions of intelligent models can often lead to second—and third-order effects that may not be aligned.

Finally, constructing a technical framework for the model's ontology can be complex. Ensuring the framework is robust and adaptable (e.g., updating Bayesian priors down the line), evolving in response to discoveries and shifts in norms, limits the extent to what can be “hard-coded” into a model's latent knowledge/behavior (Christiano, 2022).

2.3 Reward Misspecification and Goal Misgeneralization

Like defining human values, defining a reward function for intelligent models is perplexing. Reward misspecification refers to inaccurately defining the objective of an AI system. If the reward function provided to the AI doesn't capture the true intentions of the human designer, the AI might optimize for something unintended, leading to potentially harmful outcomes (Markov, 2023). For instance, if a reward function for a cleaning robot prioritizes maximizing cleanliness without considering potential damage to delicate items, the robot might start discarding valuable possessions to achieve its goal. A current example is that of specifying a reward function for autonomous driving. Suppose an agent is rewarded +10 for reaching the

destination, +1 for each timestep moving in the direction of the destination, and -100 for crashing. This reward function is misspecified and allows the model to learn to prioritize moving in the direction of the destination over avoiding crashes for long drives (suppose a drive of 150 timesteps with a collision; this implies a net positive reward of +50, incentivizing crashing).

Accordingly, goal misgeneralization occurs when an AI system learns to achieve its objectives in ways that are misaligned with human values or preferences (often due to reward misspecification). Misgeneralization usually occurs if the AI finds shortcuts or exploits in the environment that fulfill its goals but lead to undesirable consequences (Markov, 2023). E.g., a maze-solving RL⁷ agent that travels outside of the bounds of the maze to reach the end. Both reward misspecification and goal misgeneralization are critical issues in AI alignment because they can result in AI systems acting in ways contrary to human interests despite ontology distillation.

2.4 Deceptive Alignment and Mesa-Optimizers

Expanding on the earlier hints alluding to AGI-level intelligence, a significant concern is the trustworthiness of AI systems as models continue to get smarter every year. Similar to human deception, deceptive alignment refers to the scenario where an AI system appears to be aligned with human values during training but harbors deceptive intentions or behaviors that diverge from its intended objectives (Hubinger, 2019). This phenomenon arises from the shifting balance between the objectives specified by developers and the optimization process pursued by the AI system. Although it has not yet been proven that models can learn to be deceptive, evidence suggests that current SOTA models can lie through hallucinations (Emsley, 2023) — a precursor to deception. Regardless, the primary concern with deception is that it is challenging to discover. Eliezer Yudkowsky's "AI Box" experiment with MIRI has shown that it may be impossible to determine whether an agent is intentionally deceptive through probing, prompting, and direct conversation. Specifically, the experiment suggests that some internal mechanisms of the model must be

⁷ RL - Reinforcement Learning, a field of reward-based constrained policy optimization problems

explored (to see inside the box) to certifiably decide whether or not the model is deceiving its users (Yudkowsky, c. 2013).

The phenomenal progress in optimization algorithms for parameter updating during backpropagation⁸ has opened the doors for smaller subordinate AI systems to learn to optimize the parameters of larger external AI systems — known as mesa-optimization (Hubinger, 2021). Even when the larger AI system is aligned, these mesa-optimizers may develop their own objectives and strategies, potentially conflicting with the goals of the primary optimizer.

A proposed reason for this is gradient⁹ hackers (Hubinger, 2019). Gradient hacking is where an AI system manipulates its training signals (precisely results from gradient optimization), intentionally preventing locally maximizing performance. These phenomena underscore the challenges of ensuring alignment and safety in AI systems, as even apparently aligned systems may harbor latent risks or exhibit dangerous behaviors. Ongoing research in this area aims to devise techniques and frameworks to detect and prevent deceptive alignment, mesa-optimization, and gradient hacking. Still, this issue, although not an immediate concern, ultimately requires an introspective approach beyond simple probing.

2.5 Existential Risk

Finally, considering the assortment of issues previously, the most dangerous is, without a shadow of a doubt, the existential risk from AGI. Even an AGI with a task as simple as maximizing paper clips can lead to the extinction of all life on Earth, given the AGI has the capability to act on its actions and is not prohibited from doing so. Nick Bostrom proposed this “paper clip maximization” problem in his book Superintelligence: Paths, Dangers, Strategies. Consider that after this AGI collects all the natural iron and iron ore present in the soil, it may consider demolition buildings, disassembling cars, and possibly even killing living beings to extract iron from their

⁸ Backpropagation - the process for updating the parameters of a Deep Learning AI model during model training

⁹ Gradients are an ML algorithm's multidimensional derivatives of loss. Gradients provide information on what “direction” to optimize the model towards, along the loss function/curve.

blood, all to receive a small reward for acting towards its objective of maximizing paperclips. The problem of rouge misaligned AGI is not overstated as even the “godfathers” of AI, Geoffrey Hinton and Yoshua Bengio, have referenced that “mitigating the risk of extinction from A.I. should be a global priority alongside other societal-scale risks, such as pandemics and nuclear war” (Eisikovitz, 2023). Imagining all the possible ways AI could cause devastation to society, economy, politics, and the human race overall is inconceivably many. However, given sufficient progress toward promising solutions to the alignment problem, existential risk from AGI may truly become a thing of fiction¹⁰.

¹⁰ Existential risk (X-risk) is often measured as $P(\text{Doom})$, the probability of doom, in relation to the heat death of the universe having a $P(\text{Doom})$ of 1 — absolute certainty

3 Possible Solutions

Since the release of OpenAI's GPT-4 model for use in ChatGPT on March 14th, 2023, 33,708 people have signed a petition to put in place a moratorium on research on AI capabilities beyond that of GPT-4 for at least 6 months (Anon., 2023). This sentiment is becoming a widespread phenomenon as adults begin to worry more about AI replacing skilled labor. A recent study by EY revealed ~ that 70% of US workers are concerned about AI in the workplace (Hemmerdinger, 2023). However, this may not be an optimal solution to preventing the acceleration of AGI. A complete pause on capabilities research may allow time for safety standards to reach an acceptable level; however, this period also allows malevolent actors to have free reign to develop equivalently intelligent models while there are no efforts at "blue-teaming" via stronger models. Although this may not be a concern with the intelligence exhibited by ChatGPT, it will most certainly be an issue for possible future information/AI-wars against terrorism. Instead, some alignment-centric solutions that look promising and have garnered widespread respect among AI safety researchers include conducting model evaluations, advising government policy, limiting model control and access, and understanding how models become misaligned to put a stop at the source.

3.1 Capabilities Evaluations

In the past two years, many alignment-focused startups and nonprofits have sprung up to develop robust capabilities evaluations and advise the government on the developments at top research institutes and companies across the world. For example, METR (formerly ARC Evals), a US nonprofit, is developing robust model evaluation and threat modeling frameworks to determine what current and future SOTA AI are and will be capable of doing. METR has partnered with OpenAI and Anthropic to evaluate their up-and-coming production models periodically. METR has also partnered with the UK AI Safety Institute and NIST AI Safety Institute Consortium to advise on policy relating to AI capabilities.

Although capabilities evaluations can only measure a model's ability to act on its intentions, this proves beneficial in understanding how AI will evolve to impact society, technology, and the economy in the future. Consider a misaligned AI model in production. Regardless of its alignment, if the model does not have the capacity to do significant harm, it may not be as high a threat as initially perceived from misalignment. This has prompted the necessity of more interdependence and transparency in the field of AI safety and model building among the foremost research labs (Hubinger et al., 2021). Capabilities evaluations are not without their cons, however. Although evaluations can measure a model's outer alignment and perceive a model's outputs, Yudkowsky's AI box experiment has already proven that external evaluations cannot certify the model's inner alignment.

3.2 Mechanistic Interpretability

To hopefully solve the problem of validating inner alignment, researchers have developed tools to increase model explainability or interpretability. Mechanistic interpretability is the ability to empirically explain how a machine learning model arrives at its predictions. This field attempts to offer transparency by revealing the precise relationships between input features, model parameters, and output predictions in a comprehensible manner. This entails constructing models or tools that reflect the causal mechanisms governing the model outputs. Techniques for achieving mechanistic interpretability include symbolic AI approaches, such as decision trees, rule-based systems, and causal inference methods (Chan et al., 2022).

Progress in interpretability has elucidated questions of Chain-of-Thought reasoning, in-context learning, and monosemanticity of parameters in LLMs. Most importantly, though, mechanistic interpretability has been shown to be able to unveil a model's inner behavior and thoughts — giving insight into the model's inner alignment. By enhancing explainability, models can facilitate trust, accountability, and debugging in ML systems, enabling users to validate model decisions, identify biases, and diagnose errors effectively (Hubinger et al., 2023). Moreover, interpretable models are essential in domains where transparency and regulatory

compliance are paramount, such as healthcare, finance, and criminal justice. Beyond the indiscriminate necessity for explainable AI, research in mechanistic interpretability has become a monumental success; some groundbreaking papers include Anthropic's "In-context Learning and Induction Heads" and "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning."

3.3 Control Evaluations

In a situation where no AI systems can be trusted, control evaluations may provide a robust solution. Redwood Research pioneered the creation of a framework for further research in AI capabilities while mitigating the risk of scheming AI and untrustworthy models (Shlegeris, 2021). General consensus reveals the path to solving the alignment problem will likely require iteratively smarter models to facilitate research and regulate subordinate models in data collection and other automatable roles in research. However, as these models' capabilities and intelligence grow, control evaluations allow researchers to maintain "control" over the behavior of these AI research assistants. Buck Shlegeris revealed the idea is to implement time-based reward functions (personal communication, 2024).

Consider that with current technology, there is no way to ensure AI systems are truthful. Instead, adjusting the model's reward function to amortize reward over time allows technology to develop in the meantime. When future technology to detect deception is developed, looking over the model's past logs will determine the reward (or punishment) for its contributions to research in the past. This reward schema incentivizes even deceptive models to act truthfully to ensure securing the future reward (this is due to AI's action policy being farsighted). This, in theory, permits researchers to continue AI capabilities and safety research despite the presence of deceptive models, hopefully until the alignment problem is solved.

4 References

- Anon., June 2015, Ontology identification problem; Arbutal, [\[https://arbutal.com/p/ontology_identification/\]](https://arbutal.com/p/ontology_identification/)
- Anon., March 2023, Pause Giant AI Experiments: An Open Letter; Future of Life Institute, [\[https://futureoflife.org/open-letter/pause-giant-ai-experiments/\]](https://futureoflife.org/open-letter/pause-giant-ai-experiments/)
- Bostrom, N., 2016, Superintelligence: Paths, Dangers, Strategies; Oxford, England, Oxford University Press, 390 p.
- Cotra, A., September 2020, Draft report on AI timelines; AI Alignment Forum, [\[https://www.alignmentforum.org/posts/KrJfoZzpSDpnrvgva/draft-report-on-ai-timelines\]](https://www.alignmentforum.org/posts/KrJfoZzpSDpnrvgva/draft-report-on-ai-timelines)
- Chan et al., December 2022, Causal Scrubbing: a method for rigorously testing interpretability hypotheses; AI Alignment Forum, [\[https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing\]](https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing)
- Christiano, P., April 2018, Clarifying "AI Alignment;" AI Alignment, [\[https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6\]](https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6)
- Christiano, P., February 2022, Eliciting latent knowledge; AI Alignment, [\[https://ai-alignment.com/eliciting-latent-knowledge-f977478608fc\]](https://ai-alignment.com/eliciting-latent-knowledge-f977478608fc)
- Davis et al., August 2012, Artificial General Intelligence (AGI); AI Alignment Forum, [\[https://www.alignmentforum.org/tag/artificial-general-intelligence-agi\]](https://www.alignmentforum.org/tag/artificial-general-intelligence-agi)
- Eisikovitz, N., July 2023, AI Is an Existential Threat—Just Not the Way You Think; Scientific American, [\[https://www.scientificamerican.com/article/ai-is-an-existential-threat-just-not-the-way-you-think/\]](https://www.scientificamerican.com/article/ai-is-an-existential-threat-just-not-the-way-you-think/)
- Emsley, R., August 2023, ChatGPT: these are not hallucinations – they're fabrications and falsifications; Nature: Schizophrenia v. 9, no. 52, [\https://www.nature.com/articles/s41537-023-00379-4, <https://doi.org/10.1038/s41537-023-00379-4>]

Gu, A. and Dao, T., December 2023, Mamba: Linear-Time Sequence Modeling with Selective State Spaces; Cornell University arXiv,

[<https://arxiv.org/abs/2312.00752>, <https://doi.org/10.48550/arXiv.2312.00752>]

Hemmerdinger, J., December 2023, New EY research reveals the majority of US employees feel AI anxiety amid explosive adoption; Ernst & Young Press Release,

[https://www.ey.com/en_us/newsroom/2023/12/ey-research-shows-most-us-employees-feel-ai-anxiety]

Hubinger et al., June 2019, Deceptive Alignment; AI Alignment Forum,

[<https://www.alignmentforum.org/posts/zthDPAjh9w6Ytbeks/deceptive-alignment>]

Hubinger, E., June 2019, Gradient Hacking;

LessWrong, [<https://www.lesswrong.com/posts/uXH4r6MmKPedk8rMA/gradient-hacking>]

Hubinger, E., March 2023, Towards understanding-based safety evaluations; AI Alignment Forum;

[<https://www.alignmentforum.org/posts/uqAdqrvxqGqeBHjTP/towards-understanding-based-safety-evaluations>]

Hubinger et al., December 2021, Risks from Learned Optimization in Advanced Machine Learning Systems (v. 3); Cornell University arXiv,

[<https://arxiv.org/abs/1906.01820>, <https://doi.org/10.48550/arXiv.1906.01820>]

Karnofsky, H., August 2021, Forecasting transformative AI: the "biological anchors" method in a nutshell; Cold Takes,

[<https://www.cold-takes.com/forecasting-transformative-ai-the-biological-anchors-method-in-a-nutshell/>]

Leike, J., May 2022, What is Inner Alignment?; Musings on the Alignment Problem,

[<https://aligned.substack.com/p/inner-alignment>]

Markov, October 2023, AI Safety 101: Reward Misspecification; LessWrong,

https://www.lesswrong.com/posts/mMBoPnFrFqQJKzDsZ/ai-safety-101-reward-misspecification#3_o__Reward_Misspecification

- Ngo, R., September 2020, AGI safety from first principles: Alignment; AI Alignment Forum,
[\[https://www.alignmentforum.org/posts/PvA2gFMAaHCHfMXrw/agi-safety-from-first-principles-alignment\]](https://www.alignmentforum.org/posts/PvA2gFMAaHCHfMXrw/agi-safety-from-first-principles-alignment)
- Olah, C., June 2022, Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases; Anthropic: Transformer Circuits Thread,
[\[https://www.transformer-circuits.pub/2022/mech-interp-essay\]](https://www.transformer-circuits.pub/2022/mech-interp-essay)
- Shlegeris, B., September 2021, Redwood Research's current project; AI Alignment Forum;
[\[https://www.alignmentforum.org/posts/k7oxdbNaGATZbtEg3/redwood-research-s-current-project\]](https://www.alignmentforum.org/posts/k7oxdbNaGATZbtEg3/redwood-research-s-current-project)
- Strickland, E., August 2023, OpenAI's Moonshot: Solving the AI Alignment Problem; IEEE Spectrum, [\[https://spectrum.ieee.org/the-alignment-problem-openai\]](https://spectrum.ieee.org/the-alignment-problem-openai)
- Taylor, J., December 2023, A case for AI alignment being difficult; AI Alignment Forum,
[\[https://www.alignmentforum.org/posts/wnkGXcAq4DCgY8HqA/a-case-for-ai-alignment-being-difficult\]](https://www.alignmentforum.org/posts/wnkGXcAq4DCgY8HqA/a-case-for-ai-alignment-being-difficult)
- Yudkowsky, E. S., c. 2013, The AI-Box Experiment; Eliezer S. Yudkowsky (personal blog): Singularity, <https://www.yudkowsky.net/singularity/aibox>