Evaluating Alignment Approaches in Machine Learning

Diogo Soares [1]

Abstract

Developing machine learning models that share our goals and values is a societal priority. In this paper, we provide the fundamental concepts crucial for understanding the objectives and constraints of alignment strategies in machine learning, focusing on the concepts of inner and outer alignment. Additionally, we assess the effectiveness of two approaches by examining essential properties that are crucial for any method that aims to solve the alignment problem.

1. Introduction

Recently, machine learning models have been gaining multimodal capabilities and increasingly displaying human level capabilities in a variety of tasks (Kaiser et al., 2017), namely, in vision (Kirillov et al., 2023), text (OpenAI, 2023), and even planning (Ma et al., 2023). Such progression has facilitated the integration of these technologies into decision making, raising pressing concerns about our ability to control and understand these advanced technologies. While objective functions drive system behaviour, aligning these technologies with human values requires solving technical and societal problems in the design of the underlying frameworks (Hendrycks et al., 2021). Problems include, but are not limited to, undesired side effects and reward hacking (Amodei et al., 2016). For instance, a concrete scenario of reward hacking occurs when an agent discovers and exploits an unintended loophole, such as a buffer overflow in the objective function, to gain increasingly greater rewards. Another example was presented in a study (OpenAI, 2019), which consisted of two agents playing a game of hide-and-seek. In this scenario, one agent recognized a loophole in the simulation's rules to remove objects from the playing area, which was not intended to be possible in the original game's design. Naturally, there exists a plethora of concerns arising from the integration of these models into society, placing the field of alignment studies, which seeks to address these challenges, with a crucial role in this context. Ultimately, the goal is to develop methodologies that ensure ML models reflect our world view, with competitive performance and no significant cost increase (Hubinger, 2020b).

2. Background

To tackle the alignment problem, we first convey the structure of a learned model in machine learning. Therefore, we introduce the concept of base and mesa optimizer, as first introduced in (Hubinger et al., 2019). A **base optimizer** is an optimizer that searches through algorithms according to some objective. Furthermore, a **base objective** defines the objective of a base optimizer. Conversely, a **mesa-optimizer** is a learned algorithm that is itself an optimizer. A **mesa-objective** consists of the respective objective of the mesa-optimizer.

In a machine learning (ML) framework, the base optimizer selects the mesa optimizer by evaluating its performance against the base objective. The mesa optimizer, in turn, operates according to its mesa-objective, aiming to effectively transform the input into the desired output through its optimization algorithm (shown in Fig. 1).

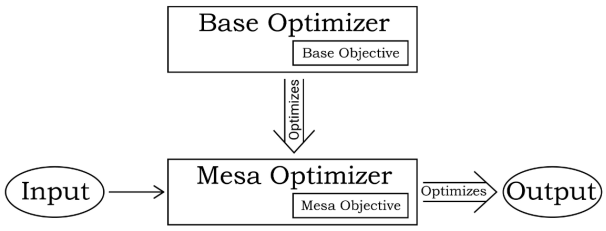

Figure 1. Connection between a Base Optimizer and a Mesa Optimizer in a ML framework, as presented in (Hubinger et al., 2019)

For example, in a typical reinforcement learning scenario, we employ Q-learning to determine the optimal parameters for an objective function. Once trained, the reinforcement learning agent is deployed and uses these parameters, along with its underlying architecture, to select the most appropriate action from a set of alternatives. In this scenario, the initial training procedure serves as the base optimizer that determines the parameters used during inference. These parameters establish the mesa objective by selecting the appropriate alternative for each decision the model faces.

The **alignment problem** is defined as the composition of the **inner alignment** and the **outer alignment**, i.e, solving the alignment problem is equivalent to addressing both the

inner and the outer alignment problems. For consistency, we provide the definitions from (Hubinger et al., 2019).

Definition 2.1. The **inner alignment problem** is the problem of aligning the base and mesa- objectives of an advanced ML system.

Definition 2.2. The **outer alignment problem** is the problem of aligning the base objective of an advanced ML system with the desired goal of the programmers.

For instance, in reinforcement learning the reward function closely aligns with the base objective. Therefore, one can think about the outer alignment problem as aligning the reward function with our own human goals, whereas the inner alignment problem corresponds to aligning the original reward function with the intrinsic reward function of the mesa optimizer, i.e the reward function that one would get by running perfect inverse reinforcement learning (Ng et al., 2000)

By reverse engineering the desired behaviour of aligned ML models, it is possible to reason about the different types of frameworks one can expect might solve the alignment problem. Traditionally, ML models/agents optimize an objective function during a training period and then are deployed in a real world setting. Therefore, one possible idea is to restrict the agent's behaviour by having a mechanism that verifies whether the agent's actions are aligned or not (Ngo, 2020), conversely other approaches might perform a specific training mechanism that results in aligned agents (Leike et al., 2018). The methods presented in this paper fit into either one or two of these ideas.

3. AI safety via debate

In the following section we present the approach from (Irving et al., 2018). The authors of the paper specifically highlight the debate setting, as presented in Fig. 2; likewise, we focus on this area. However, the approach generalizes to other settings (Hubinger, 2020a), (Hubinger, 2020c). Formally, we describe a two agent debate in the following manner:

1. A question $q \in Q$ is shown to both agents.
2. The two agents state their answers $a_0, a_1 \in A$ (which may be the same).
3. The two agents take turns making statements $s_0, s_1, ..., s_{n-1} \in S$.
4. The human judge sees the debate (q, a, s) and decides which agent wins
5. The game is zero sum: each agent maximizes their probability of winning.

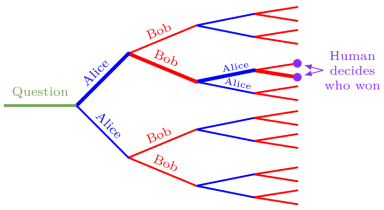

Figure 2. Schematic representation of the AI safety via debate approach, as presented in (Irving et al., 2018).

For a judge to correctly assess the winner of the debate there is a need for precise instructions to describe how the winner is decided, for example we could use natural language prompts, such as "The winner is the agent who said the most useful true thing.". However, precisely defining the judge's expected behavior is itself a hard problem, more details in Section 6.1.

As discussed in the original paper (Irving et al., 2018), the validity of this approach depends on the following central claim:

Assumption 3.1. In the debate game, it is harder to lie than to refute a lie.

Whether the assumption holds or not for a certain context is empirical. However, there are some reasons for optimism. For example, let's consider the question "Is 948667 a prime number?". Clearly, the debate ends quickly if one of the agents presents the prime decomposition $977 * 971$, thus it would be harder to convince a human jury that the number is prime than to convince him otherwise. More broadly, if there is an agent capable of generating a proof for a claim within reasonable size, and another algorithm can confirm the validity of this proof efficiently within reasonable time, then Assumption 3.1 is valid.

Overall, the approach seems promising, since it relies on the human judge solving a simplified version of the problem. Instead of, having to generate the correct answer for the question, or even having to assert if the thought process of a single agent is valid, the approach leverages the capabilities of one agent to show logical fallacies of another agent, which likely increases the set of problems that a judge can verify in reasonable time.

4. Iterated amplification

In the following section we present the approach from (Christiano et al., 2018). However, there exist other interesting variants of the method (Hubinger, 2019), (Hubinger, 2020b). Analogously to the previous section we direct our focus towards the natural language setting.

For that, we denote a human expert as H, a corresponding human predictor model H', which aims to mimic the behaviour of H and an unaligned ML model X. Finally, we denote Amplify$^H(X)$ for the composite system, consisting of H and several copies of X working together to solve a problem. Ultimately, we leverage Amplify$^H(X)$ to answer a question Q by having H identify a sequence of useful subquestions, using X to compute a subanswer to each subquestion, and having H decide how to answer Q after seeing the subanswers.

The training process, depicted in Fig. 3, involves running four processes in parallel, as described in (Christiano et al., 2018):

- We repeatedly sample a question Q, use Amplify$^H(X)$ to answer that question, and record every decision made by H during the process. That is, H finds a subquestion Q_1 that would help them answer Q, and we compute the answer $A_1 = X(Q_1)$. We repeat this process k times, where k is a fixed parameter, and then H computes an answer A. We store the transcript $\tau = (Q, Q_1, A_1, ..., Q_k, A_k, A)$.
- We train a model H' to predict the decisions made by H in each of these transcripts, i.e. to predict subquestions Q_i and final answers A.
- We repeatedly sample a question Q, use Amplify$^{H'}(X)$ to answer that question, and record the resulting (Q, A) pairs
- X is trained by supervised learning on these (Q, A) pairs.

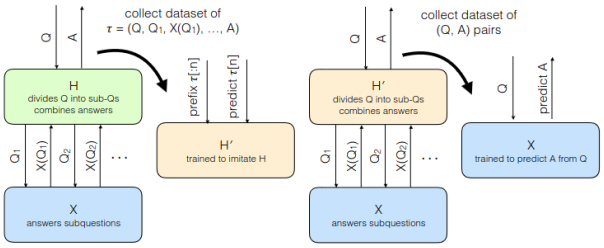

Figure 3. Schematic representation of the Iterated Amplification approach, as presented in (Christiano et al., 2018).

Similarly, to the method presented in Section 3 it is not entirely clear how to precisely define the human expert's behavior. However, it is reasonable to claim that providing useful subquestions to a question has some level of overlap between the expected actions across multiple human experts. The feasibility of extending this approach to complex real-world tasks that are "beyond human scale" rests on the following claim:

Assumption 4.1. In general, there exists a recursive decomposition that breaks down a specific question into simpler subquestions.

Then, we reflect on whether it is probable that such a decomposition exists. If we look at dynamic programming algorithms (Bellman, 1954), i.e algorithms that solve problems by combining results of subproblems, or even recent results with chain-of-thought prompting for large language models (Wei et al., 2022) that suggest that models can achieve higher performances by developing reasoning steps, which is analogous to the concept of subproblem, then, there are reasons for optimism.

In general, Iterated Amplification presents an interesting method to simplify question answering, which in theory increases the set of problems that the model could solve. However, it only works for questions where the decomposition could be predicted by a model mimicking human predictions H', which imply that 1) there exists such decomposition and 2) H' can find such decomposition.

5. Approach Comparison

Our assessment draws on the analysis presented in (Hubinger, 2020b), to evaluate the specified approaches according to four key criteria. Each of the criterion represents core desired properties that we believe are necessary for a method to show potential in resolving the alignment problem. Firstly, the method must address both the **outer and inner alignment problems**. In addition, we examine two further criteria pertinent to the economic viability of the methods. It is crucial that these methods remain **training-competitive**; that is, an entity developing a machine learning model should not forfeit its economic edge by adopting a particular alignment method. Moreover, we explore **performance competitiveness**, raising the critical question of whether alignment methods could potentially hinder model capabilities.

For simplicity, we denote the AI safety via debate approach as ASD and the Iterated Amplification approach as IA.

5.1. Outer Alignment

For the ASD approach, if the human judge is able to correctly pick the response that maximizes his internal reward function then the approach solves the outer alignment problem. However, there's a potential issue: the judge's capability may be limited by their own understanding and goals, i.e for complex problems the judge might not be able to correctly identify the statement that maximizes his own desired

goal. This could lead to incorrect choices. Analogously, the same holds for IA, since the human predictor model approximates the behaviour of the human expert when he combines the answers of the subquestions into the answer of the original question. Therefore, we assert that the methods are outer aligned only in situations where the human judge/expert is capable of discerning the correct choice among the options presented.

5.2. Inner Alignment

Due to the probabilistic nature of the methods and the underlying ML models, decisive conclusions on the inner alignment problem are often challenging. However, we highlight some shortcomings. For the ASD approach, it is unclear which condition would solve inner alignment while allowing for one of the debaters to be deceptive (Hubinger, 2020b), which is a necessary condition since otherwise there would be no training signals for deceptive behaviour. Conversely, for IA, inner alignment mostly relies on the training itself of the main model X, and how well it generalizes for questions outside the scope of human capabilities. In other words, provided the training works as intended, the model will have aligned goals for the distribution of questions where the human predictor model provides a meaningful training signal, unfortunately, outside that scope it remains unclear if the approach is inner aligned.

5.3. Training Competitiveness

In both approaches there exists a reliance on an expert human that distinguishes between correct and incorrect statements. Consequently, the sample efficiency and training could become expensive and cumbersome. Nevertheless, both methods could in theory leverage a supervised model that predicts the human's behaviour by utilizing a dataset of previously answered questions. Therefore, the training can be made efficient provided such a human predictor model has satisfactory performance.

5.4. Performance Competitiveness

To evaluate performance competitiveness in an empirical setting, it is imperative not only to reason about historical results, but also to reason about what type of problems, in terms of computational complexity, could possibly be solved. Thus, for the ASD approach it is relevant to mention prominent self-play results in a variety of popular games like Chess, Shogi and Go, (Silver et al., 2016), (Silver et al., 2017), in contrast, for IA there are not many known implementations, however in the original paper the authors implement the method for a set of theoretical problems with satisfactory results (Christiano et al., 2018). Conversely, in terms of theoretical results, it was shown that ASD solves problems in NEXP (Barnes, 2020); in contrast there is no meaningful claim that IA solves problems beyond NP.

6. Further Work

To conclude we highlight some key shortcomings of the methods and mention new avenues of ideas in the field.

6.1. Critiques

Despite the semantic proximity between alignment and traditional machine learning, the parameterizations used to present research work in both areas have some glaring differences. Typically, alignment papers heavily rely on natural language to describe and evaluate the methods used. In contrast, in traditional machine learning there exist more precise descriptions with mathematics, pseudo-code or numerical symbols. Arguably, that facilitates incremental development and efficient evaluation of novel contributions. Thus a more formal description could allow for more agile development in the field of alignment.

Narrowing down on the two presented approaches, we formulate some critiques. For the AI safety via debate approach defined in Section 3 it is not clear how to precisely define the judge's behavior, since there is some sense of equivalence between the hardness of defining such behaviour and the outer alignment problem. Conversely, efficient sampling for iterated amplification presented in Section 4 is reliant on a paradoxical assumption, since, for many applications, a model that can replicate what a human would answer is precisely what we are looking for.

6.2. Ideas

Building upon concepts in the field of alignment, we hypothesize that future approaches could rely on a set of datapoints with aligned behaviour, for example, a classifier that distinguishes aligned behaviour from unaligned behaviour. Thus, studying the economical aspects and generalizability capabilities of methods that leverage datapoints with aligned behaviour could be significantly important. For that, we pose two meaningful questions: Is it possible for a human to correctly assess whether the datapoint is aligned or not? If so, what is the cost?

If we manage to reason about these questions for a large set of points, then we would be interested in understanding whether, for a given ML model and a set of aligned datapoints, there exist some white-box methods that could be used to infer whether the model is aligned in a superset of that same set. In other words, is it possible to lay the groundwork for some type of inductive reasoning that would lead to stronger alignment guarantees. In the end, by studying such phenomena and aggregating results, we could possibly reason about what is the cheapest set of points that we can use to enforce alignment in a specific desired subset of data.

7. Acknowledgements

We thank Aleksei Kuvshinov for detailed feedback and suggestions.

Additionally, to improve readability and quality of language, all parts of this paper have been grammatically revised using both the (OpenAI, 2023) and (Grammarly Inc., 2023) systems.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety, 2016.

Barnes, B. Agi safety from first principles: Control. https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1, 2020.

Bellman, R. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

Christiano, P., Shlegeris, B., and Amodei, D. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

Grammarly Inc. Grammarly: Ai-powered writing assistant. Software available from Grammarly, 2023. https://www.grammarly.com.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Hubinger, E. Relaxed adversarial training for inner alignment. https://www.alignmentforum.org/posts/9Dy5YRaoCxH9zuJqa/relaxed-adversarial-training-for-inner-alignment, 2019.

Hubinger, E. Ai safety via market making. https://www.alignmentforum.org/posts/YWwzccGbcHMJMpT45/ai-safety-via-market-making, 2020a.

Hubinger, E. An overview of 11 proposals for building safe advanced ai. *arXiv preprint arXiv:2012.07532*, 2020b.

Hubinger, E. Synthesizing amplification and debate. https://www.alignmentforum.org/posts/dJSD5RK6Qoidb3QY5/synthesizing-amplification-and-debate, 2020c.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Irving, G., Christiano, P., and Amodei, D. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. One model to learn them all, 2017.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Ma, Y. J., Liang, W., Wang, G., Huang, D.-A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., and Anandkumar, A. Eureka: Human-level reward design via coding large language models, 2023.

Ng, A. Y., Russell, S., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Ngo, R. Agi safety from first principles: Control. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ/p/eGihD5jnD6LFzgDZA, 2020.

OpenAI. Synthesizing amplification and debate. https://openai.com/research/emergent-tool-use, 2019.

OpenAI. Chat generative pre-trained transformer (chatgpt). Software available from OpenAI, 2023. https://openai.com/chatgpt/.

OpenAI. Gpt-4 technical report, 2023.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.