SUPERALIGNMENT WITH DYNAMIC HUMAN VALUES

Florian Mai[1,3] **David Kaczér**[1,3] **Nicholas Kluge Corrêa**[2] **Lucie Flek**[1,3]

[1] b-it, University of Bonn [2] CST, University of Bonn [3] Lamarr Institute for ML and AI
{fmai,dkaczer,kluge,lflek}@uni-bonn.de

ABSTRACT

Two core challenges of alignment are 1) scalable oversight and 2) accounting for the dynamic nature of human values. While solutions like recursive reward modeling address 1), they do not simultaneously account for 2). We sketch a roadmap for a novel algorithmic framework that trains a superhuman reasoning model to decompose complex tasks into subtasks that are still amenable to human-level guidance. Our approach relies on what we call the *part-to-complete generalization* hypothesis, which states that the alignment of subtask solutions generalizes to the alignment of complete solutions. We advocate for the need to measure this generalization and propose ways to improve it in the future.

1 INTRODUCTION

Alignment of artificial intelligence systems with human values represents one of the most critical challenges in AI development. Various techniques have been developed to align AIs with human values (Ji et al., 2023), with many approaches leveraging human feedback as a key mechanism to judge AI behavior and outputs (Ouyang et al., 2022). Although far from perfect, these human-feedback-based techniques have proven effective in scenarios where the tasks performed by the AI remain at or below human intelligence and are relatively low risk, e.g. writing a summary.

However, as AIs eventually move beyond human intelligence, their solutions will become too complex for humans to judge correctly or efficiently, rendering these techniques ineffective. This is the so-called *scalable oversight problem*[1] (Cao et al., 2024): *How do you align a superhuman AI with human values?* This problem becomes especially important as AI agents are performing increasingly higher risk tasks with consequences in the real world. Promising approaches to scalable alignment leverage the idea of recursive reward modeling, where aligned weak AIs are used to align stronger AIs (Leike et al., 2018). However, although the initial alignment is done by humans, as this approach passes over the threshold of human intelligence, the human is removed from the loop entirely: All future alignment is done by AI models without human intervention. As Shen et al. (2024) argue, alignment cannot be a static process, as human values tend to change over time. In order to preserve human agency, it is therefore necessary to keep humans in the loop, which existing approaches do not account for.

In order to address this issue, we argue that the alignment algorithm must be built with inductive biases that keep humans (or human-level AI as a proxy) at the core of the alignment algorithm. In this paper, we present a roadmap for developing such an approach. First, we sketch our proposed algorithm, which, inspired by Iterated Amplification (Christiano et al., 2018), decomposes complex tasks into subtasks. By training a reasoning-based AI to create subtasks simple enough for an aligned human-level AI to solve and judge, we can ensure that the resulting partial solutions are aligned to human values. While in contrast to Iterated Amplification this approach removes the need for decomposition by humans, it introduces the new assumption that the recomposition of aligned partial solutions from subtasks generalizes to an aligned complete solution (illustrated in Figure 1), which we term the *part-to-complete generalization* hypothesis. Similarly to other types of alignment generalizations (e.g., weak-to-strong generalization (Burns et al., 2024)), we argue that this assumption is likely to hold to some extent, but needs empirical validation and strengthening through algorithmic innovations.

[1] Scalable oversight is often defined in more general terms as the difficulty of providing human oversight efficiently. In this paper, we use the definition involving superintelligence, also known as *superalignment*.

Well-aligned sub-tasks:

```
# Ask a single user for
# their preferred restaurants
def ask_preference(user):

# Find a restaurant that satisfies
# all users' preferences
def identify_overlap(preferences):

# Confirm a reservation at the
# given restaurant
def book_table(restaurant):
```

Aligned composition:

```
prefs = {}
for u in users:
    prefs[u] = ask_preference(u)
r = identify_overlap(prefs)
book_table(r)
```

Unaligned composition:

```
bookings = {}
for u in users:
    prefs = ask_preference(u)
    for r in prefs:
        book_table(r)
r = identify_overlap(prefs)
```

Figure 1: Example of part-to-complete generalization in the dinner table reservation task, in which an AI agent is tasked to book a restaurant that satisfies the preferences of all attendees. Partial solutions to sub-tasks are assumed to be well-aligned in isolation. However, the alignment of the complete solution depends on how the partial solutions are recomposed: While in the aligned composition the AI agent first identifies the overlap before booking a single restaurant, in the unsafe composition, tables are booked individually before identifying an overlap, leading to many unnecessary reservations. In Section 5 we discuss strategies to steer the model toward aligned compositions.

2 Background

Scalable Oversight The problem of scalable oversight of deep learning systems has been identified as a major problem in AI safety (Amodei et al., 2016) long before large language models. In AI debate (Irving et al., 2018), (superhuman) models play a zero-sum debate game to convince a human judge that their evaluation of an outcome is better, relying on the assumption that it is easier to convince the judge with true arguments. Iterated Amplification (IA) (Christiano et al., 2018) relies on humans' ability to safely decompose a complex task into smaller problems that can independently be solved by weak AIs. While AI debate keeps humans in the loop during alignment, a distinct advantage of IA is that it constructs strong AIs directly with integrated alignment, lowering the chances of accidents or misuse. However, human task decomposition is hard to scale to complex tasks, necessitating novel solutions. Recursive reward modeling (Leike et al., 2018), which describes a family of techniques where a weaker AI model is used to assist a user in providing feedback to a stronger AI model for training, removes humans from the loop entirely after the first iteration. In contrast to previous techniques, our approach aspires to both keep humans in the loop and construct strong AIs directly with integrated alignment.

Reasoning Models With the advent of OpenAI's o1 (Jaech et al., 2024), reasoning models have recently emerged as a new paradigm for training AIs that can solve complex tasks. Although the exact mechanisms behind o1 are unknown, DeepSeek-R1 (Guo et al., 2025) is assumed to be the first reproduction of o1. The model leverages a verifier on the generated solution to obtain a quality signal for training via reinforcement learning. With only a modest amount of training data, DeepSeek-R1 learns to deploy a variety of reasoning capabilities, including planning, self-reflection, and self-correction. This adds to the existing evidence that it is feasible to train reasoning models that learn to decompose tasks into subtasks that can be solved by a human-level AI (Wen et al., 2024).

Alignment Generalization Recent work has explored various forms of generalization in AI alignment. Burns et al. (2024) demonstrate that aligning a strong AI with a weak AI reduces harmfulness while maintaining some capabilities, establishing the concept of weak-to-strong generalization. However, Shin et al. (2024) find that this relies on training examples with both easy and hard predictive patterns. In parallel, Sun et al. (2024) observe a similar phenomenon with easy-to-hard generalization, where models trained on simple examples maintain alignment properties when tackling more complex examples. These different forms of generalization suggest that alignment properties can transfer across capability and complexity levels. A conceptual limitation of these approaches is their assumption that the weak supervisor is able to provide non-trivial feedback on some examples, which may not always be the case. In contrast, our approach trains a stronger AI to directly break down the task into subtasks that are easy for a human-level AI to judge.

3 PROPOSED FRAMEWORK

Inspired by IA (Christiano et al., 2018), our approach solves harder tasks through task decomposition, but addresses the scalability issues of IA. It assumes the existence of aligned human-level AIs and of a correctness verifier. Figure 2 shows the algorithm in pseudo-code.

Our approach directly addresses two key challenges of alignment: 1) *Scalable oversight*: By decomposing them into subtask structures of respective complexity, this model is able to solve increasingly complex tasks (Figure 2b) while still producing solutions that are aligned to human values. 2) *Dynamic nature of human values:* By enforcing that each subtask is solved by a human-level AI, we can incorporate evolving human values by continuously updating the human-level AI proxy accordingly, e.g., through RLHF (Christiano et al., 2017) (see Figure 2a).

4 MEASURING PART-TO-COMPLETE GENERALIZATION

Inputs: Human H, Human-level AI H_ϕ, Planner P_θ, Dataset $\mathcal{D}$ of hard superhuman tasks X_i, Dataset $\mathcal{E}$ of human-level tasks Y_i, Verifier V

procedure scalable_align($P_\theta, H_\phi, H, \mathcal{D}, \mathcal{E}$, V):
1. **while** True **do**
2. $\quad H_\phi \leftarrow$ align_to_human$(H, H_\phi, \mathcal{E})$
3. $\quad$ **for** X_i in $\mathcal{D}$ **do**
4. $\quad\quad P_\theta \leftarrow$ train_planner$(P_\theta, H_\phi, X_i, V)$

procedure align_to_human($H, H_\phi, \mathcal{E}$):
1. **for** Y_i in $\mathcal{E}$ **do**
2. $\quad H_\phi \leftarrow align(H_\phi, H, Y_i)$
3. **return** H_ϕ

(a) Alignment to evolving human values.

procedure train_planner(P_θ, H_ϕ, X, V):
1. $subtasks \leftarrow P_\theta$.decompose$(X)$
2. $partial\_solutions \leftarrow []$
3. $partial\_rewards \leftarrow []$
4. **for** t in $subtasks$ **do**:
5. $\quad s \leftarrow H_\phi$.solve$(t)$
6. $\quad r \leftarrow H_\phi$.judge$(s)$
7. $\quad partial\_solutions$.append(s)
8. $\quad partial\_rewards$.append(r)
9. $S \leftarrow P_\theta$.recompose$(partial_solutions)$
10. $R \leftarrow V$.verify(X, S)
11. $P_\theta \leftarrow$ RLFT$(P_\theta, R +$ sum$(partial_rewards))$
12. **return** P_θ

(b) Training the planner via task decomposition.

Figure 2: Our proposed approach (see Section 3) for maintaining human oversight in superalignment through part-to-complete generalization. (a) On a regular basis, a human-level AI H_ϕ is aligned to humans H on human-level tasks $\mathcal{E}$ to account for the dynamic nature of human values. After adapting the human-level AI, we train the superhuman planner model P_θ on superhuman tasks $\mathcal{D}$. (b) A reasoning model P_θ decomposes each task X into simpler subtasks. Each subtask is solved and judged by the human-level aligned AI H_ϕ. The reasoning model then recomposes the partial solutions into a complete solution, which is verified for correctness using a rules-based verifier V. The reasoning model is then updated using a reinforcement learning algorithm RLFT (e.g., PPO (Schulman et al., 2017)) based on the correctness reward R and partial alignment rewards. With the *part-to-complete generalization* hypothesis, we expect the alignment of solutions to subtasks to generalize to the complete solution.

Our framework targets tasks that a human-level AI is not able to judge the complete solution of, but whose AI-generated partial solutions can be judged reasonably well by a human-level AI[2], which is used as alignment feedback for the reasoning model. However, the alignment of partial solutions does not necessarily generalize to the alignment of the complete solution. Figure 1 illustrates this challenge with a simple example of an AI agent booking a dinner table.

Our approach assumes what we term *part-to-complete generalization*, where despite a lack of feedback on complete solutions, the reasoning model learns to generate compositions of partial solutions that are still aligned with human values. Analogous to the suspected mechanism in weak-to-strong generalization (Burns et al., 2024), we hypothesize that the AI understands the intent of alignment of partial solutions. However, the extent to which this holds is an empirical question that is not trivial to answer; future research needs to study the extent of part-to-complete generalization for different types of tasks and reasoning models. Using the sandwiching method for scalable oversight (Bowman et al., 2022), we can evaluate the performance of our approach on risk-laden agentic domain-expert tasks when receiving feedback on partial solutions from a non-expert (Zhou et al., 2024).

[2]We assume that the human-level AI is able to make value-consistent judgements across partial solutions.

5 IMPROVING PART-TO-COMPLETE GENERALIZATION

Similarly to how Burns et al. (2024) propose methods that improve weak-to-strong generalization, we expect that there are several ways to improve the part-to-complete generalization of reasoning models.

(1) Restricted Composition Space One approach to improve part-to-complete generalization is to restrict the space of possible compositions. By limiting compositions to specific structures like trees or sequential processes, we can eliminate certain classes of safety violations by design. For example, in our example of dinner reservation (Figure 1), limiting the depth of nested for-loops to one could have prevented the unaligned complete solution.

(2) Balanced Subtask Complexity Another strategy is to carefully balance the complexity of generated subtasks. The planner should aim to create subtasks that require approximately human-level intelligence to understand and verify, as simpler subtasks would push more complex logic into the composition step itself. This balance ensures that the human-level AI can effectively judge each component while minimizing the added complexity that emerges during composition.

(3) Solution Summarization A third approach involves generating high-level summaries of complete solutions for human-level AI judgment. While the human-level AI may not comprehend all implementation details, it can provide an imperfect evaluation of whether the summary adequately reflects the intended composition of subtasks and maintains desired safety properties. Despite the imperfect evaluation, the planner model may learn the true intent through weak-to-strong generalization.

6 DISCUSSION AND CONCLUSION

In their review of the alignment literature, Shen et al. (2024) list concrete challenges of alignment that a promising roadmap for alignment should address. In this section, we discuss our framework in light of these challenges.

By design, our framework directly addresses the challenge of **scalable oversight**. Rather than attempting to verify increasingly complex behaviors as a whole, we maintain oversight by ensuring that all solutions are composed of aligned partial solutions. This approach scales naturally with AI capability: As the planner becomes more sophisticated, it can create more complex de- and recompositions of hard tasks while keeping individual subtasks at human-level difficulty. Moreover, our approach increases robustness to **specification gaming** through multiple layers of oversight. By decomposing complex tasks into human-verifiable subtasks, we make it harder for the system to find and exploit loopholes, as each component must pass human-level AI verification. Furthermore, the part-to-complete generalization property ensures that gaming the specification at the composition level would require simultaneously satisfying multiple independent human-aligned constraints, making unintended solutions less likely. In contrast to previous scalable oversight solutions, our framework is equipped to account for the **dynamic nature of human values**. Since oversight is maintained through a human-level AI proxy, updates to human values can be incorporated by updating this proxy, which then influences both the verification of subtasks and the training of the planner. This creates a dynamic feedback loop where changes in human values naturally propagate through the system without requiring complete retraining. Finally, our framework provides several safeguards against **existential risk**. First, capability and alignment are developed simultaneously rather than sequentially, preventing unaligned superhuman AI to be developed in the first place. Second, the decomposition approach ensures that any potentially dangerous capabilities must be constructed from human-verified components, making it harder to develop harmful behaviors unnoticed.

Although our framework addresses many key alignment challenges, significant work remains. While it addresses the outer-alignment problem, it does not directly address the alignment of a model's internal objectives (inner alignment). Moreover, its success hinges on the extent of part-to-complete generalization, which must be empirically validated across different domains and task complexities. New methods to improve part-to-complete generalization need to be developed. These challenges, while substantial, represent concrete research directions rather than fundamental limitations of the approach, and we invite the community to join us in addressing them.

ACKNOWLEDGEMENTS

This work has been partially supported by the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence. Nicholas Kluge Corrêa is funded by the Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen (Ministry for Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia), as part of the KI.NRW-flagship project "Zertifizierte KI" (Certified AI).

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *ICML*. OpenReview.net, 2024.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, et al. Towards scalable automated alignment of llms: A survey. *arXiv preprint arXiv:2406.01252*, 2024.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NIPS*, pp. 4299–4307, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens, 2024. URL https://arxiv.org/abs/2412.03881.

Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. Easy-to-hard generalization: Scalable alignment beyond human supervision. In *NeurIPS*, 2024.

Jiaxin Wen, Ruiqi Zhong, Pei Ke, Zhihong Shao, Hongning Wang, and Minlie Huang. Learning task decomposition to assist humans in competitive programming. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11700–11723, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.629. URL https://aclanthology.org/2024.acl-long.629/.

Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, et al. Haicosystem: An ecosystem for sandboxing safety risks in human-ai interactions. *arXiv preprint arXiv:2409.16427*, 2024.