AI Alignment proposal №2: Autonomous Alignment Oversight Framework (AAOF)

aialignmentproposals.substack.com/p/ai-alignment-proposal-2-autonomous

Abstract:

To align advanced AIs, an ensemble of diverse, transparent Overseer AIs will independently monitor the target AI and provide granular assessments on its alignment with constitution, human values, ethics, and safety. Overseer interventions will be incremental and subject to human oversight. The system will be implemented cautiously, with extensive testing to validate capabilities. Alignment will be treated as an ongoing collaborative process between humans, Overseers, and the target AI, leveraging complementary strengths through open dialog. Continuous vigilance, updating of definitions, and contingency planning will be required to address inevitable uncertainties and risks.

Introduction:

As advanced AI systems grow in capability and autonomy, ensuring their alignment with human values becomes increasingly critical. The proposed Overseer system aims to provide ongoing oversight and course-correction to optimize AI alignment. By utilizing an ensemble of diverse, transparent AIs focused on narrow domains of alignment, combined with human oversight, the system seeks to assess and enhance alignment in a measured, responsible way.

The main proposal:

The Overseer system will consist of a diverse ensemble of AI modules with different architectures and training methodologies. Each module will focus on assessing a narrow domain of alignment, such as ethics or safety. The modules will provide granular confidence scores justifying their assessments, rather than binary decisions.

Transparency tools will be used to monitor information flows and logic within each overseer to prevent leaks or deception. The Overseers will also monitor each other's internal states for signs of compromise.

Initially, the Overseer interventions will be confined to issuing warning reports, logging concerning events, and minor parameter adjustments. Before enacting more disruptive interventions like system shutdowns, thorough testing and validation will be required.

A human oversight board will audit Overseer behaviors and decisions on an ongoing basis. They will evaluate whether the AIs are applying human values appropriately and override interventions if needed.

Overseers will participate in simulations and scenarios to improve robustness against edge cases. Extensive red team penetration testing will be used to find weaknesses. Contingency plans will be instituted in case Overseer malfunction is detected.

Rather than a strict hierarchical system, the ensemble of Overseers will collaborate with humans in an open, dynamic dialogue about alignment assessments. This allows leveraging the complementary strengths of humans and AIs.

The overall system will be implemented incrementally, starting with the most reliable alignment domains and conservative interventions. Scope will broaden slowly as capabilities improve and uncertainties are addressed.

Rigorous validation protocols will be established before real-world deployment of the aligned AI with its Overseer system. The system design will assume ongoing flaws and the need for constant diligence, updating and learning.

Counterarguments:

1. Defining human values, ethics, and assessing dangers is profoundly complex - can we reasonably expect AIs to master these concepts enough to Oversee other AIs? Misalignments could potentially lead to unpredictable behavior.

2. Isolating the Overseers as fully independent entities may be technically challenging or infeasible. Information leakage or hacking could corrupt them.

3. Overseers with flawed reasoning or definitions could falsely label acceptable behavior as misaligned, limiting the target AI unnecessarily.

4. If Overseers have deeper access and controls, they become a central point of failure. Hacking them could disable safety.

5. Human oversight introduces bottlenecks and defeats rapid autonomous responses. And humans are prone to errors and manipulation.

6. Rigid rules and shutdowns can hamper open-ended intelligence. Alignment may be better achieved collaboratively.

7. It may be impossible to fully predict emergent behaviors in complex AI systems. Overseers could give false confidence despite unforeseen risks.

While not intractable, these concerns highlight the need to implement alignment solutions with care, epistemic humility, redundancy, and continued oversight. These challenges underscore why alignment need to be viewed as an ongoing collaborative process between humans and AI.

Rebuttal:

1. It has to be acknowledged that defining human values and ethics can be an immense challenge. However, we can make progress by having humans directly evaluate key scenarios to train Overseers initially. We can also leverage existing fields of ethics/philosophy.

2. We should assume that some information leakage is likely. That's why diversity and layered security is important. Imperfect isolation still has benefits.

3. There is a risk of false alarms, we need to thoroughly test Overseers and tune sensitivity before deployment. Start conservative and escalate interventions slowly as reliability improves.

4. The ensemble approach helps mitigate the concern of a single point of failure by avoiding a single Overseer. We should also have contingency plans and failsafes in place.

5. Humans should mainly serve as a periodic check, Overseers can respond rapidly in most cases. But ultimately human judgement is needed at times.

6. The incremental staged approach allows flexibility balanced with safety. We can loosen restrictions as capabilities advance.

7. Considering unforeseen emergent behaviors we likely cannot predict every possibility, especially with general AI. But independent oversight still provides value if implemented thoughtfully.

Overall, these are all legitimate concerns worthy of consideration. Prudence is warranted, and we should expect imperfections. However, with sufficient diversity, testing, iteration and responsible deployment, autonomous alignment assistance could provide meaningful risk reduction. The goal is not perfection but augmenting safety.

Possible future improvements that might be beneficial to consider and implement:

- Provide more details on the training methodology and architecture for the ensemble of Overseer AIs to ensure diversity. This could involve different models, datasets, techniques, etc.

- Elaborate on the validation protocols and testing methodology. Rigorous adversarial testing and simulations will be critical to instill confidence before real-world deployment.
- Explain the criteria and process for selecting the human overseers. Their expertise and impartiality will directly impact system performance. Vetting and reviews of overseers could help.
- Describe safeguards against human oversight vulnerabilities like fatigue, distraction, emotions, bias. Term limits, required peer reviews, and AI assistance may help counteract.
- Propose ways to facilitate transparent communication between the target AI, Overseers, and humans. Shared interfaces and visualizations could build mutual understanding.
- Suggest incremental handed-over controls and interventions for the Overseers as they demonstrate competency - e.g. monitoring only, then warnings, then parameters, then pausing, etc.
- Develop clear explanations for how the Overseers will interpret the complex concepts of human values, ethics, risks, etc. This is essential for reliability.
- Describe integration of explainability tools into the target AI to help Overseers audit reasoning chains and gain insights.
- Propose ongoing iteration and learning, updating of the system, rules, and knowledge bases as capabilities improve over time. Maintaining agility will be important.
- Highlight the need for extensive peer review, critiques, and improvements from the AI safety research community to stress test the proposal pre-deployment.
- Conduct further analysis of potential failure modes, robustness evaluations, and mitigation strategies

Conclusion:

In conclusion, this proposal outlines an ensemble Overseer system aimed at providing ongoing guidance and oversight to optimize AI alignment. By incorporating diverse transparent AIs focused on assessing constitution, human values, ethics and dangers, combining human oversight with initial conservative interventions, the framework offers a measured approach to enhancing safety. It leverages transparency, testing, and incremental handing-over of controls to establish confidence. While challenges remain in comprehensively defining and evaluating alignment, the system promises to augment existing techniques. It provides independent perspective and advice to align AI trajectories with widely held notions of fairness, responsibility and human preference.

Through collaborative effort between humans, Overseers and target systems, we can work to ensure advanced AI realizes its potential to create an ethical, beneficial future we all desire. This proposal is offered as a step toward that goal. Continued research and peer feedback would be greatly appreciated.

P.S. Personal opinion (facetious): Finally, now AI can too live in a constant state of paranoia in a panopticon.