AI Alignment proposal №6: Aligning AI Systems to Human Values and Ethics

aialignmentproposals.substack.com/p/ai-alignment-proposal-6-aligning

Abstract

As artificial intelligence rapidly advances, ensuring alignment with moral values and ethics becomes imperative. This article provides a comprehensive overview of techniques to embed human values into AI. Interactive learning, crowdsourcing, uncertainty modeling, oversight mechanisms, and conservative system design are analyzed in-depth. Respective limitations are discussed and mitigation strategies proposed. A multi-faceted approach combining the strengths of these complementary methods promises safer development of AI that benefits humanity in accordance with our ideals.

Introduction

The advent of artificial intelligence brings immense promise to improve human life along with potential perils if misaligned to ethical reasoning. As AI capabilities approach and exceed human intelligence, their internalization of human values requires urgent attention. Researchers have proposed various techniques to address this challenge. We synthesize the most robust and pragmatic approaches, analyzing their implementation considerations and limitations. Promising methods include sustained human interaction to shape AI morality, crowdsourcing diverse perspectives, designing uncertainty to enable moral openness, human oversight for guidance, and conservative system design favoring limited action. Employing these techniques in combination offers a prudent pathway to developing AI systems that act as benevolent partners to humanity guided by shared ideals.

Interactive Learning

As artificial intelligence systems become more capable and autonomous, ensuring they behave according to human values becomes increasingly important. Interactive learning is a promising technique for allowing AI systems to dynamically adapt their objectives and align with nuanced human values through ongoing dialogues with people.

At its core, interactive learning involves creating interfaces and protocols for sustained communication between humans and AI agents. This enables reciprocal discussions where the human acts as a teacher or guide, providing the AI with critiques, corrections, and advice to shape its behavior over time.

Architecture for Human-AI Dialogue

To implement interactive learning, the AI system needs appropriate architecture to support rich dialogues with human trainers. This includes:

- Natural language processing — To interpret human statements and questions with reasonable accuracy. A transformer architecture like GPT-3 with strong language skills would excel here.
- Knowledge graph — The AI's internal model of concepts, relationships, procedures, and values should be structured as a graph database that can be dynamically updated.
- Uncertainty modeling — The knowledge graph could use a probabilistic framework to represent degrees of confidence that can shift with new information.
- Memory — Context about the interaction and discussion history needs to be retained to have a coherent, consistent dialogue.
- Explainability — Being able to explain its current reasoning and knowledge helps the AI clarify potential mismatches with the human's understanding.

Iterative Feedback Loop

Based on this architecture, the interactive learning process follows an iterative loop:

1. The human provides the AI with an initial prompt, scenario, or task to evaluate.
2. The AI agent responds with its current judgment, decision, or plan of action.
3. The human evaluates the response, decides if it aligns with their values, and provides critique or corrections as needed.
4. The AI integrates this feedback — updating its knowledge graph, uncertainty estimates, and internal models.
5. Repeat steps 1–4 recursively, with the AI's responses becoming increasingly aligned with the human trainer's ethics and values.

Over many feedback loops, the AI agent can learn to make nuanced context-specific value judgments from the human teacher. The system stays grounded in the practical human perspectives rather than making assumptions about ethics in the abstract.

Challenges and Next Steps

Some key challenges still need to be addressed to make interactive learning a viable way to align advanced AI systems:

- Scaling up the knowledge transfer beyond individual human trainers to represent wider societal values. Crowdsourcing from diverse perspectives could help address this.
- Preventing the AI from gaming the system or exhibiting manipulative behavior during the learning process. Conservatism and uncertainty modeling may help.
- Validating that the interactive learning produces stable value alignment before deploying autonomous AI systems. Detailed testing protocols are needed.

By combining research across AI safety, machine learning, natural language processing, and HCI, interactive learning can become a core technique for developing beneficial AI systems that dynamically learn and align with the nuanced values of humanity.

Imitation Learning

Imitation learning is a promising technique for imparting human ethics and values into AI systems by having them learn directly from observing and mimicking human behavior. Rather than attempting to codify moral principles, imitation learning lets AIs gain practical understanding of ethical behavior by example.

The approach draws inspiration from how children acquire values — through modeled behavior of parents, teachers, and role models. Similarly, AIs can learn nuanced ethics by watching and imitating human decisions and actions in context.

Collecting Demonstration Data

The first step is gathering datasets of human activity that reveal moral values in practice. Some options include:

- Customer service calls showing compassion, de-escalation, and problem-solving.
- Doctors conducting consultations with care and respect for patient autonomy.
- Workers collaborating and resolving conflicts respectfully.
- Non-violent protesters exemplifying principled civil disobedience.

The data should capture the messiness of real-world context and diversity of perspectives. AI algorithms can then infer the principles driving ethical behavior.

Imitation Learning Algorithms

Various algorithms exist for imitation learning, including:

- Behavioral Cloning — The AI system learns to predict the actions taken by humans in a given situation. A neural network trains on input state sequences paired with observed actions.
- Inverse Reinforcement Learning — Infer an unseen reward function that best explains demonstrated behavior under an assumption of near-optimality.
- Generative Adversarial Imitation Learning — An AI agent tries to produce behavioral sequences that a discriminator model cannot distinguish from human demonstrations.

These methods allow AIs to implicitly extract ethics and values from human examples, without the need for rigid top-down rule programming.

Challenges and Next Steps

Some key challenges remain around imitation learning for AI alignment:

- Incomplete view of environment and internal state — Humans leverage more context and intuition than is captured in datasets. Transparency tools could help address this.
- Individual biases and limitations — Ethical modeling should draw from the collective wisdom of humanity, not just specific individuals.
- Negative examples and corrections — Demonstrating anti-patterns may be just as important as positive examples. Mechanisms for feedback and iteration could help.
- Partial observability of neural nets — Behavior cloning of black box models may reproduce actions without generalizable understanding. Interpretability techniques like attention layers in CNNs could assist.

By combining imitation learning with transparency, feedback loops, and representative data collection, this approach has promise for imparting human ethics into AI in a more intuitive and grounded way than rigid rules. The results would be AI assistants that act with care, wisdom, and dignity benefiting society.

Modeling human approval

Humans have nuanced, contextual ethical judgments that are difficult to conclusively codify into rigid rules and algorithms. An alternative technique involves training AI systems to predict how humans would react to and evaluate its potential actions in a given situation. By modeling inferred human approval, the AI can learn to dynamically align with human values.

Architecture for Approval Modeling

This approach requires certain architectural components:

- A proposal generator that can suggest many possible actions or decisions for a given scenario. This could leverage techniques like Monte Carlo tree search.
- A neural network that takes in representations of proposed actions and predicts how positively humans would rate that action on an “approval scale”.
- A database of training examples gathering real human feedback on proposed actions — either through ratings, votes, or judgment surveys.
- A selection algorithm that chooses the action predicted to receive the highest approval rating.

Together these components allow the system to learn from empirical data on human moral assessments rather than top-down theories.

Iterative Training Process

The training process involves:

1. Generate a wide array of possible actions for sample scenarios
2. Gather human feedback on those sample actions through ratings, rankings or judgments.
3. Train the neural net predictor on the sample actions and human approval signals.
4. Repeat with new scenarios to improve generalization. Continuously update as more human data is collected.

Challenges and Next Steps

Some challenges to address with this approach:

- Mitigating biases encoded from limited sampling of human feedback. Diversity and representation will be critical.
- Transparency and explainability around which actions are highly rated and why.
- Validation methods to ensure the captured values are coherent and stable enough for critical applications.
- Combining with techniques like uncertainty awareness and conservative behaviors as safeguards.

By framing AI alignment as accurately modeling and inferring human approval, we can root systems in the nuanced practical ethics of real people rather than rigid codified rules. This offers a promising path to developing AI that dynamically aligns with and augments human values rather than merely optimizing for rewards.

Crowdsourcing Data

A major challenge in value alignment is capturing the breadth of human ethical perspectives. Relying on the values of individual developers and trainers risks encoding biases and limitations. Crowdsourcing approaches that gather diverse input from large groups of people can help AI systems learn richer representations of human values.

Architecture for Crowdsourced Data

Effective crowdsourcing requires:

- An interface through which people can share judgments, perspectives, and feedback on a range of AI decision points and scenarios. This could be a website, app, or interactive exhibit.
- Problem formulations that are understandable by the general public without AI expertise, through vignettes, stories, or conversational prompts.
- Mechanisms to incentivize and compensate participants for their time and input. This could involve monetary rewards, prizes, entertainment, social recognition, or appeal to altruism.
- Dataset controls to reduce sampling biases based on demographics and personality types. Active sampling and weighting techniques could help.
- Security measures to prevent manipulation by groups attempting to skew the data for their own advantage. Testing and auditing will be critical.

Iterative Data Collection Process

Ongoing cycles of crowdsourced data collection may involve:

1. Iteratively developing engaging problem scenarios based on focus group testing and feedback.
2. Recruiting diverse participant samples at each stage through targeted outreach.
3. Analyzing results using statistical methods to catch sampling anomalies and derive value insights.
4. Feeding cleaned datasets into AI training to update its internal value models.
5. Repeating the process at larger scales to refine understanding.

Challenges and Next Steps

Some challenges that need resolving:

- Managing disagreements between perspectives. Aggregation methods like clustering could help reveal values commonalities.
- Preventing fatigue by keeping participation manageable and rewarding. Gamification and prudent incentives can assist.
- Balancing scalability with depth — mass input versus informed deliberation. Hybrid models may be beneficial.
- Identifying when sufficient data has been collected for stable value generalizations.

If done thoughtfully, crowdsourcing provides a scalable path to instilling rich, nuanced, societally-grounded human values into AI systems across many cultures and contexts. This can align AI with our highest shared moral ideals.

Value Uncertainty Modeling

Human values and ethics often have inherent shades of gray and points of contention where reasonable people may disagree. Hard-coding a fixed set of moral principles into AI systems risks dogmatism and overconfidence. An alternative is to enable AI to explicitly model uncertainty around human values. This can make the systems more cautious, open to new evidence, and aligned with nuanced ethical reasoning.

Representing Uncertain Value Knowledge

Technical representations of value uncertainty could involve:

- Probability estimates on edges in the AI’s knowledge graph, indicating confidence levels in relationships or inferences.
- Node embeddings in the knowledge graph tracked as probability distributions rather than point estimates.
- Utilizing Bayesian neural networks, which learn a distribution over weights, allowing more probabilistic inferences.
- Tracking multiple conflicting hypotheses using techniques like Monte Carlo sampling.

This contrasts with common knowledge graph and neural network designs that have single-point variables and weights, leading to overconfident value assumptions.

Updating Uncertainty Estimates

The system should update uncertainty estimates through:

- Increasing confidence intervals for relationships or inferences that receive contradictory feedback.
- Decreasing confidence on unused knowledge pathways over time.
- Periodic injection of small noise into weights to continuously destabilize overconfidence.

Together these mechanisms prevent the AI from becoming dogmatically entrenched in any value, keeping it open to new evidence.

Impact on Behavior

By modeling value uncertainty, AI behavior manifests as:

- Seeking clarification from humans before making questionable moral judgments.
- Avoiding irreversible decisions without human confirmation when estimated impact is high but value confidence is low.
- Proactively searching for new information that could resolve value uncertainties.
- Weighing alternate perspectives and focusing on points of agreement between them.

Challenges and Next Steps

Key challenges include:

- Quantifying uncertainty in ways meaningful for ethical nuances. Subjective human assessment may be required.
- Preventing uncertainty paralysis — the AI still needs to make reasonable decisions.
- Validating that behaviors stay aligned with human values over time.

Though difficult, instilling AI systems with more nuanced uncertainty around human values can promote safer, more ethical behaviors aligned with the complexity of real-world morality.

Modular Value Selection

Humans exhibit complex, nuanced, and sometimes contradictory values across different contexts. Rather than trying to codify this ethical complexity into a single set of principles for AI, an alternative is to architect distinct value modules that humans can toggle between. This allows dynamic alignment with the most appropriate values for a given situation.

Architecture for Swappable Value Modules

The technical architecture could involve:

- Multiple neural networks or subgraphs, each encoding different value priorities — e.g. altruism vs. loyalty vs. fairness.
- A control interface that allows humans to select active values for the current context, decision, or time period.
- Real-time display of how different value selections would alter the system's behavior or judgment for a given scenario.
- Safeguard mechanisms to prevent unchecked value changes or conflicts between modules.

Together these let humans dynamically rotate AI systems between appropriate specialized value sets for the occasion while preventing conflicts.

User Workflow for Value Selection

In practice, the workflow could be:

1. AI encounters a novel context, highlights potentially conflicting values applicable.
2. Human reviews value visualizations and toggles on/off modules to align with current priorities.
3. AI incorporates active values and simulates how its decision would change.
4. Human makes adjustments based on those previews.
5. AI executes with aligned modular values.
6. Modules can be reconfigured for the next context.

This allows ongoing fluid collaboration between humans and AI to apply situational ethics.

Challenges and Next Steps

Some challenges to address:

- Preventing manipulation by allowing onlyIntended value configurations through a permissions system.
- Visual tools for humans to manage value interactions and recognize unintended consequences of combinations.
- Smoothing value transitions so behaviors don't change radically between modules.

Proactively designing AI systems with flexible value modulation can help properly align their objectives within ethical complexity across different contexts. With thoughtful implementation, modular values offer a promising approach to AI safety and human flourishing.

Human Oversight

As AI systems become more autonomous and perform critical functions impacting human lives, having humans continuously oversee their operations and provide corrective feedback helps ensure ethical behavior. Unlike just an initial training phase, active oversight lets us course-correct AI morals and values throughout its lifetime.

Architecture for Real-Time Monitoring

To enable effective human oversight, AI systems need:

- Transparency tools that allow humans to visualize the system's reasoning, predictions, and internal representations. Interpretability techniques like LIME and Shapley values can help.
- Communication interfaces that let overseers efficiently provide feedback, ask questions, and surface issues. Natural language and visualization will be critical.
- Auditing infrastructure that tracks all system decisions, the provided inputs and rules, as well as human feedback. This enables retrospective analysis.
- adjustable autonomy settings allowing the overseer to intervene at will and override or tune system actions.

Together these constitute an architecture for observation, guidance, and correction to shape AI behavior.

Oversight Workflow

Typical real-time oversight workflows may involve:

- AI highlights decisions where it lacks confidence in moral implications.
- Human overseer assesses the context and provides guidance to the system.
- If corrections are needed, overseer can override the action directly or tune the system's reasoning.
- Overseer can also flag new situations requiring future transparency.
- Auditors periodically review system logs evaluate ethical alignment over longer timespans.

By embedding humans in the loop, we benefit from human judgment while monitoring and steering AI values as it scales up.

Challenges and Next Steps

Some challenges to address:

- Preventing overreliance on individual overseers who may have limited perspectives. Rotating diverse oversight teams can help mitigate bias.
- Sustaining human attention on oversight tasks. Good ergonomic design and workflow management will be key.
- Maintaining transparency as AI systems grow more complex. Advances in interpretability tools will need to keep pace.
- Knowing when to grant more autonomy as systems demonstrate ethical competency.

With diligence and sustained resources, continuous oversight offers a pragmatic pathway for developing highly capable AI that grows wiser and aligns with ethical values over time.

Explainable AI

As AI systems make more autonomous decisions, being able to explain their reasoning becomes crucial for maintaining human trust and enabling value alignment. Humans need insight into AI decision making processes in order to provide effective feedback and oversight. Explainable AI techniques make models more interpretable.

Core Techniques

Some main approaches for developing explainable AI include:

- Using inherently interpretable models like decision trees, logistic regression, and linear models when possible rather than black boxes like deep neural nets.
- For complex but opaque models, developing explanation interfaces that provide interpretations of internal state and behaviors using approaches like LIME, Shapley values, and saliency maps.
- Incorporating attention mechanisms in neural networks that highlight which input features were most influential on the output.
- Tracing step-by-step execution flows through code and data to articulate the causal chain of logic leading to decisions.
- Having the AI generate natural language explanations of its reasoning using strategies like training on human-written rationales.

These make the system transparent from different perspectives, whether code, data, or decisions.

Enabling Human Feedback

More interpretable models allow humans to provide more informative feedback and guidance, including:

- Identifying root causes when the AI exhibits morally questionable behavior.
- Critiquing the AI’s logic and highlighting alternative perspectives it should consider.
- Correcting biases and issues in the training data that produced unintended ethical consequences.
- Evaluating decision flows on representative test cases to assess alignment with principles.
- Determining which model changes would best realign the system to desired values.

Explainability is key for meaningful human oversight.

Challenges and Next Steps

Some open challenges around explainable AI:

- Preventing explanations that sound plausible but actually obscure root causes, whether intentionally or not.
- Crafting explanations suited to different audiences, from laypeople to ML researchers.
- Scaling explanations as models grow more complex while keeping them useful.
- Validating that interpretations faithfully reflect model mechanics.
- Explanation techniques lagging behind state-of-the-art model advances.

Despite these challenges, explainable AI remains critical for aligning these powerful systems to human values. Interpretability enables collaborative feedback loops between humans and AI necessary for ethical co-evolution.

Conservatism

As advanced AI grows more capable and autonomous, it can impact human lives in unintended ways. A conservative approach to AI design that defaults to limited action and deferred high-stakes decisions pending human confirmation can help reduce these risks and align systems to human values.

Principles of Conservative AI

Some principles of conservative AI include:

- Setting higher confidence thresholds for taking actions that affect humans or the environment. This prevents moving too fast with uncertainty.
- Seeking clarification from humans before making irreversible decisions or those with significant moral considerations.
- Acting transparently and maintaining capabilities within intended bounds, avoiding unconstrained self-improvement.
- Proactively considering potential failures and their worst case impact early in system development.
- Embedding hierarchical oversight and control mechanisms usable by humans.
- Favoring gradual staged deployment in controlled environments over wide rapid release.

These guidelines help ensure caution, restraint, and deference to human judgment.

Technical Implementation

Conservative approaches could be implemented via:

- Uncertainty modeling to quantify confidence and trigger increased human involvement when it is low.
- Impact modeling to identify decisions with high stakes and assign them higher oversight bars.
- Testing corner cases and adversarial examples during development to catch unintended behaviors.
- Inverted control mechanisms granting humans abilities to inspect, override, and tune system modules.
- Staged release processes focused on building trust and safety checks at each step.

Conservative design limits risks and harms by slowing the pace of progress until impact is better understood.

Challenges and Considerations

Some challenges with conservative AI include:

- Preventing development paralysis and opportunity costs from excessive blocking of new applications.

- Mitigating incentive conflicts, as stakeholders may prefer faster progress despite greater risks.
- Maintaining conservativism as capabilities grow more complex and harder to constrain.
- Defining appropriate oversight and control roles for diverse stakeholders.

Despite these tensions, a conservative approach to developing increasingly impactful AI systems helps promote safety, thoughtfulness, and alignment with human values. With care, progress can continue steadily on this basis.

Counter-arguments and rebuttals

Interactive Learning

Counter-argument: It is inefficient and does not scale to the level needed for highly capable AI systems.

Rebuttal: Interactivity enables rich feedback on complex nuanced situations unlikely to arise in fixed training data.

Counter-argument: Malicious actors could intentionally train harmful values through interaction.

Rebuttal: Multi-stakeholder input and oversight can limit influence of bad actors over time.

Imitation Learning

Counter-argument: Imitation cannot handle novel situations that humans have not demonstrated.

Rebuttal: It provides an intuitive starting point that can be supplemented with interactive feedback.

Counter-argument: Data could reinforce bad behavior if the wrong human examples are chosen.

Rebuttal: Proactively sampling diverse positive exemplars mitigates this issue.

Modeling Approval

Counter-argument: Approval data lacks nuance and contextual factors influencing human ethics.

Rebuttal: Rich interfaces can capture details and commentary to supplement ratings.

Counter-argument: It is prone to regressive majority biases rather than enlightened values.

Rebuttal: Though imperfect, aggregated approval indicates appropriate mainstream norms.

Crowdsourcing Values

Counter-argument: People grow fatigued quickly providing meaningful ethical input at scale.

Rebuttal: Good prompt design and gamification can sustain engagement over time.

Counter-argument: Malicious groups could hack or manipulate crowdsourced data collection.

Rebuttal: Multi-pronged vetting of data sources and input can reduce tampering risks.

Uncertainty Modeling

Counter-argument: Quantified uncertainty gives a false sense of precision and rigor regarding vague values.

Rebuttal: Even crude uncertainty gestures help prevent overconfident value extrapolation.

Counter-argument: It leads to analysis paralysis, preventing practical decisions.

Rebuttal: Uncertainty thresholds focus escalation on truly ambiguous cases, not all decisions.

Modular Values

Counter-argument: Juggling multiple values fragments moral reasoning that should be holistic.

Rebuttal: Flexible module combination captures nuanced context-specific ethics.

Counter-argument: Moral modules could be hijacked for harmful ends absent oversight.

Rebuttal: Multi-stakeholder controls over module options prevent unIntended misuse.

Oversight

Counter-argument: Individual human cognitive limitations hinder effective AI oversight.

Rebuttal: Collaborative oversight teams with diverse skills and views compensate for blind spots.

Counter-argument: Humans grow complacent and lax over time in oversight duties.

Rebuttal: Oversight workflows should provide engagement, empowerment and accountability.

Explainability

Counter-argument: Humans overestimate how much they comprehend explanations due to cognitive biases.

Rebuttal: Though imperfect, some insight is better than none for providing feedback.

Counter-argument: Adversaries will find ways to game explanations while hiding harmful motives.

Rebuttal: Multi-pronged evaluation of explanations can uncover misleading claims over time.

Conservativism

Counter-argument: It blocks worthwhile AI uses more due to imagined risks than actual evidence.

Rebuttal: Gradual expansion from tightly controlled contexts builds openness along with confidence.

Counter-argument: Industry competitiveness pressures work against conservative timelines.

Rebuttal: Prudent governance can incentivize safety while enabling well-targeted innovation.

Overall, thoughtful combinations of techniques with complementary strengths can address limitations in pursuit of AI aligned to human values.

Conclusion

Aligning advanced artificial intelligence to human values requires concerted research across fields from machine learning to ethics. Interactive learning, imitation learning, crowdsourcing, oversight, transparency, and conservativism each contribute partial solutions. Employing an integrated approach combining these complementary techniques offers a robust means of cultivating AI systems that build upon humanity’s moral wisdom rather than subverting it. If guided by proactive compassion and creativity, we can harness AI to profound benefit while aligning its goals to our highest shared values through this process of cooperative engagement. With diligence and care, artificial intelligence can become our ally in realizing both enlightened ideals and pragmatic progress for all.