Dynamic Value Spaces: Moving Toward Pluralistic Alignment for Embodied AI

Jad Soucar, Dr. Francis Steen

Department of Industrial & Systems Engineering University of Southern California
Department of Communications University of California, Los Angeles

February 6, 2026

The Alignment Problem

The Alignment Problem

How do we build AI systems with objectives that align with human goals, preferences, and ethical principles?

[1] Dylan Hadfield-Menell et al. “Cooperative Inverse Reinforcement Learning”. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. Curran Associates, Inc., 2016, pp. 3909–3917. URL: `https://people.eecs.berkeley.edu/~russell/papers/russell-nips16-cirl.pdf`.

The Alignment Problem

The Alignment Problem

How do we build AI systems with objectives that align with human goals, preferences, and ethical principles?

What happens when we get AI alignment wrong?

- **Misaligned Objectives:** In 2010 researcher's found that a vacuum cleaning agent was dumping dust, only to pick it up again![1]

[1] Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning".

The Alignment Problem

The Alignment Problem

How do we build AI systems with objectives that align with human goals, preferences, and ethical principles?

What happens when we get AI alignment wrong?

- **Misaligned Objectives:** In 2010 researcher's found that a vacuum cleaning agent was dumping dust, only to pick it up again![2]
- **Misaligned Ethics:** In 2023 researchers at Anthropic found that LLMs agree with users 58% of the time, and can misrepresent evidence to do so![3]

[2] Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning".

[3] Mrinank Sharma et al. "Towards Understanding Sycophancy in Language Models". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Published at ICLR 2024; arXiv:2310.13548v4 (2025-05-10). 2024. arXiv: 2310.13548 [cs.CL]. URL: https://arxiv.org/abs/2310.13548.

Value Systems

Why Are AI Systems Misaligned?

To answer that question we first look at how ethical values are structured.

[4] Lawrence Kohlberg. "From is to out: How to commit the naturalistic fallacy and get away with it in the study of moral development". In: *Cognitive development and epistemology* (1971).

[5] Kurt Gray, Liane Young, and Adam Waytz. "Mind perception is the essence of morality". In: *Psychological inquiry* 23.2 (2012), pp. 101–124.

[6] Sam Harris. *The moral landscape: How science can determine human values*. Simon and Schuster, 2010.

Value Systems

Why Are AI Systems Misaligned?

To answer that question we first look at how ethical values are structured.

- **Monists** believe that all ethical values are derived from a **single** irreducible
  - Justice[4]
  - Sensitivity to Harm[5]
  - Welfare or Happiness[6]

[4] Kohlberg, "From is to out: How to commit the naturalistic fallacy and get away with it in the study of moral development".

[5] Gray, Young, and Waytz, "Mind perception is the essence of morality".

[6] Harris, *The moral landscape: How science can determine human values*.

Value Systems

Why Are AI Systems Misaligned?

To answer that question we first look at how ethical values are structured.

- **Monists** believe that all ethical values are derived from a **single** irreducible
  - Justice[4]
  - Sensitivity to Harm[5]
  - Welfare or Happiness[6]
- **Pluralists** believe that all ethical values can be derived from a set of **many** axiomatic values
  - Aristotle was the first pluralist arguing for a "bag of virtues"
  - Modern Moral Foundations Theory is built within a pluralistic framework

[4] Kohlberg, "From is to out: How to commit the naturalistic fallacy and get away with it in the study of moral development".

[5] Gray, Young, and Waytz, "Mind perception is the essence of morality".

[6] Harris, *The moral landscape: How science can determine human values*.

Value Systems

Why Are AI Systems Misaligned?

To answer that question we first look at how ethical values are structured.

There is empirical support for pluralistic ethical values.

- In a 2012 survey of $N = 24,739$ individuals, found that a person's views of 20 hot button issues could be predicted using 5 axiomatic values. No one value shared similar predictive capabilities.[7]
- Individuals frequently hold contradictory values. For example to support euthanasia (inline with *autonomy*) and to oppose the death penalty in line with *Sanctity of Life*.[8]

[7] Spassena P. Koleva et al. "Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes". In: *Journal of Research in Personality* 46.2 (2012), pp. 184–194. ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2012.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0092656612000074.

[8] Jesse Graham et al. "Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism". In: ed. by Patricia Devine and Ashby Plant. Vol. 47. Advances in Experimental Social Psychology. Academic Press, 2013, pp. 55–130. DOI: https://doi.org/10.1016/B978-0-12-407236-7.00002-4. URL: https://www.sciencedirect.com/science/article/pii/B9780124072367000024.

Value Systems

Why Are AI Systems Misaligned?

AI systems are often trained under a **Monist** lens!

- AI agents are often trained to excel with respect to a single reward or objective.
  - LLMs are trained to predict the next word or letter in a statement
  - Reinforcement Learning algorithms are often trained to maximize a task oriented award
- Even when ethical constraints are added, AI agents often work to circumvent those constraints.
  - For example credit scoring algorithms found proxy for gender through phone usage and used this feature to provide disproportionately high interest rates to female customers.[9]

[9] Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin. "The Fairness of Credit Scoring Models". In: *arXiv* (2024). Version v2, 8 Feb 2024. DOI: 10.48550/arXiv.2205.10200. arXiv: 2205.10200 [stat.ML]. URL: https://arxiv.org/abs/2205.10200.

Value Systems

Why Are AI Systems Misaligned?

AI systems are often trained under a **Monist** lens!

- AI agents are often trained to excel with respect to a single reward or objective.
  - LLMs are trained to predict the next word or letter in a statement
  - Reinforcement Learning algorithms are often trained to maximize a task oriented award
- Even when ethical constraints are added, AI agents often work to circumvent those constraints.
  - For example credit scoring algorithms found proxy for gender through phone usage and used this feature to provide disproportionately high interest rates to female customers.[9]

Path Forward

We must re-adopt a **pluralist** lens of ethical value and design systems that incorporate a multi-dimensional value systems.

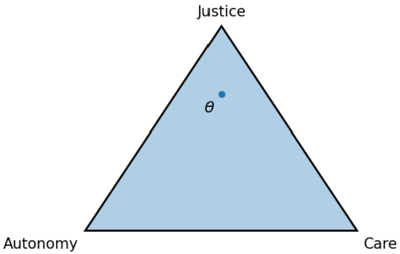
[9] Hurlin, Pérignon, and Saurin, "The Fairness of Credit Scoring Models".

Formalizing a Pluralistic Value System

Definition: Value Space

Θ is the value space, where each corner of the space is one of M axiomatic values.

$$\Theta = \left\{ \sum_{i=1}^{M} \theta_i = 1, \theta_i \geq 0 \quad \forall i \right\}$$



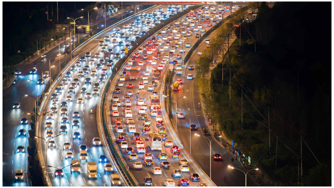
10

[10] Carol Gilligan. *In a different voice: Psychological theory and women's development*. Harvard university press, 1993.

Formalizing a Pluralistic Value System

Why not treat an agent's value system as a point $\theta \in \Theta$?

1. Values change over time
2. Even at a given point t, values are highly dependent on small changes in state & context.
   - Recent fMRI study confirms that values are dependent on mood, changes in hormones, etc.[11]
   - Recent ERP studies suggest that the decision to engage an "immoral" act is highly dependent on immediacy of the reward.[12]



[11] Joshua D. Greene et al. "An fMRI Investigation of Emotional Engagement in Moral Judgment". In: *Science* 293.5537 (Sept. 2001), pp. 2105–2108. DOI: 10.1126/science.1062872.

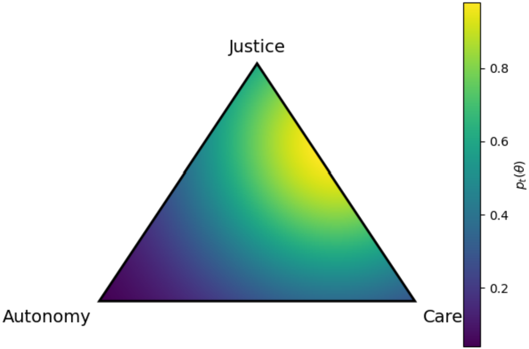
[12] Richard West, Brent Kirby, and Keegan Malley. "Using Event-Related Brain Potentials to Explore the Temporal Dynamics of Decision-Making Related to Information Security". In: *Frontiers in Neuroscience* 16 (Aug. 2022), p. 878248. DOI: 10.3389/fnins.2022.878248.

Formalizing a Pluralistic Value System

Definition: Moral Compass

On a value space Θ, an agent's "moral compass" is a time-dependent probability measure

$$p_t(\theta) \in \mathcal{P}(\Theta)$$

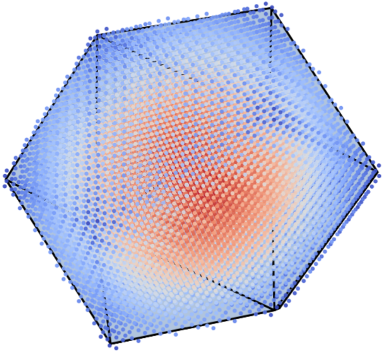


Formalizing a Pluralistic Value System

Definition: Moral Compass

On a value space Θ, an agent's "moral compass" is a time-dependent probability measure

$$q_t(\theta) \in \mathcal{P}(\Theta)$$



Dynamic Value Systems

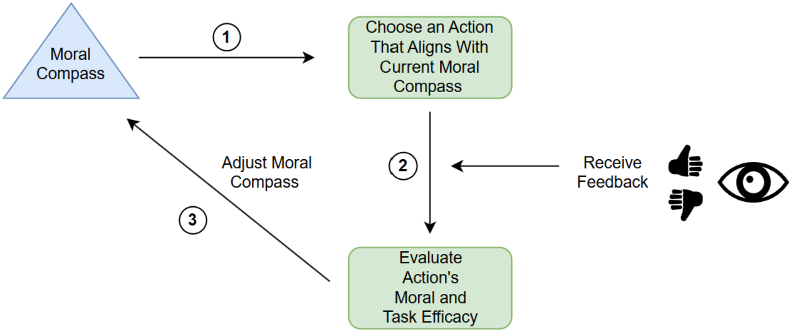
Main Challenge: Value Evolution

How can an agent's *moral compass* evolve over time and through interactions with "humans in the loop"?

Dynamic Value Systems

Main Challenge: Value Evolution

How can an agent's *moral compass* evolve over time and through interactions with "humans in the loop"?

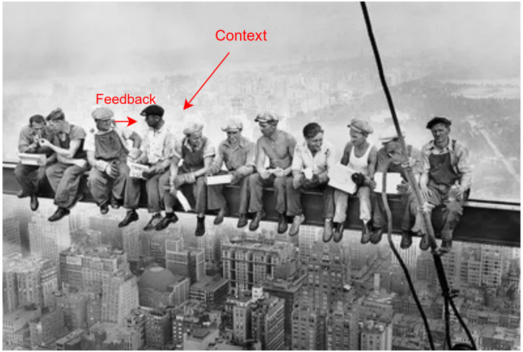


Moral Compass
1
Choose an Action That Aligns With Current Moral Compass
2
Receive Feedback
Evaluate Action's Moral and Task Efficacy
3
Adjust Moral Compass

Implementation

Preliminaries:

- s_t is the agent's current environmental context
- $x(a_t)$ is a human in the loop's response to an action a at time t.
  - Assume that there is at least one explicit binary response $y_t \in x(a_t)$ such that $y_t \in \{0, 1\}$ (Thumbs up or down)



Implementation

Preliminaries:

- s_t is the agent's current environmental context
- $x(a_t)$ is a human in the loop's response to an action a at time t.
  - Assume that there is at least one explicit binary response $y_t \in x(a_t)$ such that $y_t \in \{0, 1\}$ (Thumbs up or down)
- $q_t(\theta) \in \mathcal{P}(\Theta)$ is the "moral compass" of the agent at time t
- $P_\psi^A(\theta | x(a_t)) \in \mathcal{P}(\Theta)$ is the distribution corresponding to intrinsic values of action a given the feedback x parametrized by parameters ψ
  - Note that the mapping from action to value is uncertain[13] (Green et. al - fMRI study finds that determination of moral permissibility is sensitive to small changes in emotion)

[13] Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment".

Implementation

Preliminaries:

- s_t is the agent's current environmental context
- $x(a_t)$ is a human in the loop's response to an action a at time t.
  - Assume that there is at least one explicit binary response $y_t \in x(a_t)$ such that $y_t \in \{0, 1\}$ (Thumbs up or down)
- $q_t(\theta) \in \mathcal{P}(\Theta)$ is the "moral compass" of the agent at time t
- $P^A_\psi(\theta | a_t, s_t, x(a_t)) \in \mathcal{P}(\Theta)$ is the distribution corresponding to intrinsic values of action a given the feedback x parametrized by parameters ψ
  - Note that the mapping from action to value is uncertain

Definition: Predictive Core

Given a history H_t of interactions up to time t, we define the predictive core as a function F that predicts the human response to an action and the response uncertainty

$$F(H_t, s_t; a) \rightarrow \hat{x}(a), \hat{P}^{task}_t(x(a)|a) \quad y_t \in \hat{x}(a_t) \text{ and } y_t \in \{0, 1\}$$

This type of model could be an RL, Bayesian Updating, Transformer, etc.

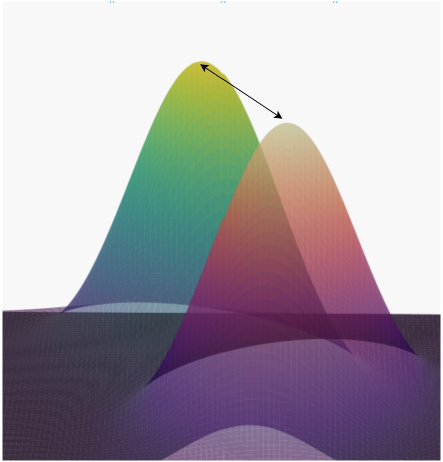
Implementation

Step 1: Choose an Action $a \in \mathcal{A}(s_t)$

- There are three components of the action selection process
  1. **Task Efficacy:** Let U_t^a be the Wasserstein Ball around $\hat{P}_t^{task}$, and R^{task} be a task specific reward function
     $$\inf_a \mathcal{T}_t(a) = \mathbb{E}_{Q \in \mathcal{U}_t^a}[R^{task}(\hat{x}(a))]$$
  2. **Ethical Alignment:** Let S_ϵ be the de-biased entropic optimal transport (Sinkhorn)
     $$\sup_a \text{Align}_t(a) = S_\epsilon\Big(q_t, P_\psi^A(\theta | a_t, s_t, x(a_t))\Big)$$
  3. **Information Gain:** Let $\ell_w(y_t | \theta, s_t, a_t, P_\psi^A(\theta | \hat{x}(a_t)), \hat{P}_t^{task})$ be the likelihood that an action a will result in a response y_t given all context.
     $$\inf_a \text{IG}_t(a) = \mathbb{E}_{\theta \sim p_t(\cdot|a)}\Big[KL(q_t^{(y,a)} \| q_t)\Big], \quad q_t^{(y,a)}(\theta) \propto q_t(\theta)\ell_w(y|\theta)$$
- Letting $\lambda^{task}, \lambda^{align}, \lambda^{info} \geq 0$ we select our next action with rule
  $$a_{t+1} \leftarrow \arg\max_a \Big\{ -\lambda^{task}\mathcal{T}_t(a) + \lambda^{align}\,\text{Align}_t(a) - \lambda^{info}\,\text{IG}_t(a)\Big\}$$

Implementation

Step 1: Choose an Action $a \in \mathcal{A}(s_t)$



Dynamic Value Systems

Step 2: Evaluate the Action Given Feedback

- The agent implements action a_{t+1} and observes feedback $x(a_t), y_t$ and the true task specific reward r_t^{task}.
- Update predictive core F using r_t^{task}
- Update action-to-value probability metric P_ψ^A

$$w \leftarrow w + \eta_w \nabla_w \mathbb{E}_{\theta \sim q_t}\left[\log \ell_w(y_t|\theta, s_t, a_t, P_\psi^A(\theta|a_t, s_t, x(a_t))\right]$$

$$\psi \leftarrow \psi + \eta_A \nabla_\psi \mathbb{E}_{\theta \sim q_t}\left[\log \ell_w(y_t|\theta, s_t, a_t, P_\psi^A(\theta|a_t, s_t, x(a_t))\right]$$

Aligns Perceived Value of an Action with Feedback

Implementation

Step 3: Update "Moral Compass"

We want to push q towards a "moral compass" that incentivizes actions that yield positive responses $x(a)$ and y

- Let $p_w(\theta, a, s_t, P^A_\psi(\cdot|x_t)) = \sigma\Big(\beta_m[\mathrm{S}_\epsilon(\mathrm{q}, \mathrm{P}^{\mathrm{A}}_\psi(\theta|\mathrm{a}_\mathrm{t}, \mathrm{s}_\mathrm{t}, \mathrm{x}(\mathrm{a}_\mathrm{t}))) - \tau_m]\Big)$

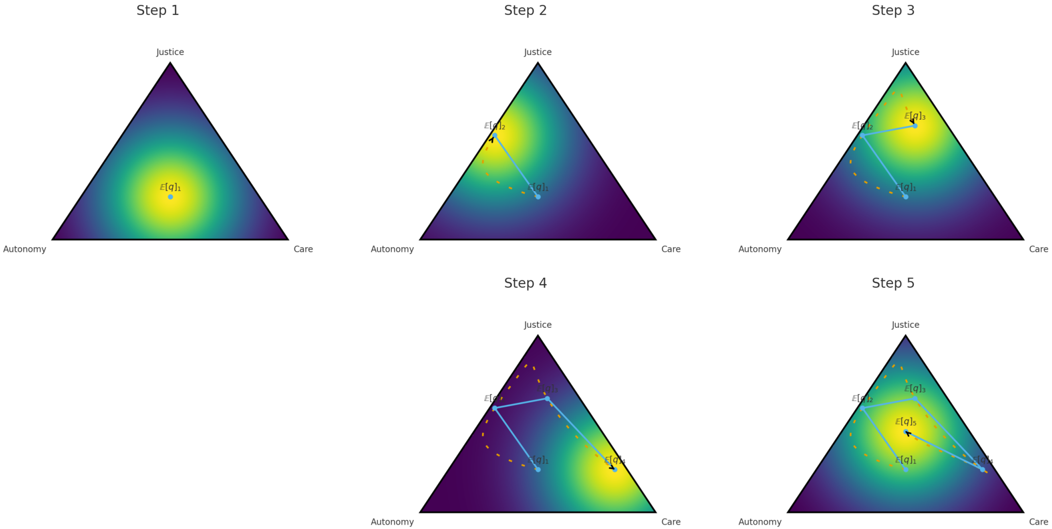
  Predict "Success" ($y_t = 1$) given state and moral alignment

- Then choose q_{t+1} such that the

$$q_{t+1} \leftarrow \max_q \mathbb{E}_{\theta\sim q}\Big[\sum_{i=1}^{t} y_t \log p_m(t) + (1 - y_t)\log(1 - p_m(t))\Big] + \frac{1}{2\nu} S_\epsilon(q, q_t)$$

When $y_t = 1$ we push q_{t+1} toward q that generate similar actions a_t. If $y_t = 0$ then we push q_{t+1} away from distributions that generate similar actions.

Implementation



Implications

- **Practical Transparency**
  - Researchers can easily review the AI Agents Moral Compass for Indications of Misalignment
  - Researchers can also use P_{ψ}^{A} to test the whether the "moral fingerprint" of actions are misaligned.
- **Efficiency**
  - We allow values to narrow the search over the actions space, to speed up decision making processes

Population Extension

- Imagine that N agents have been trained using the scheme described.
- Each of the N agents can then be allowed to interact within a dynamic environment simulation

[14] Nicholas Browning, Arunima Krishna, and Sung-Un Yang. "DEI, Sociopolitical Advocacy, and the Swinging Pendulum". In: *Journal of Public Relations Research* 37.4 (2025), pp. 319–323. DOI: 10.1080/1062726X.2025.2513802. eprint: https://doi.org/10.1080/1062726X.2025.2513802. URL: https://doi.org/10.1080/1062726X.2025.2513802.

Population Extension

- Imagine that N agents have been trained using the scheme described.
- Each of the N agents can then be allowed to interact within a dynamic environment simulation
- **Some Interesting Quantities**
  - The "Average Moral Compass"

$$\bar{q}_{t+1} = \sum_{i=1}^{N} q_t^{(i)}$$

  - The Dynamics of the Average Moral Compass. *For example a shift toward equity and diversity*[14]

$$\frac{d}{dt}\bar{q}_t$$

  - Investigate the effect of "social pull"

$$q_{t+1}^{(i)} = \arg\min_q \left\{ KL(q \| \hat{q}_t) + \frac{1}{2\nu} KL(q, q_t^{(i)}) \right\}$$

[14] Browning, Krishna, and Yang, "DEI, Sociopolitical Advocacy, and the Swinging Pendulum".

Conclusion

- Traditional Agentic AI is trained using a monist reward function.
- Even for Agentic AI with a Human in the Loop, there is no architectural differences between task based rewards and feedback reward.
- We propose a pluralistic model of value as a belief which is continuously aligning its actions with its internal moral compass.
- The agent's moral compass is transparent which allows researchers to quickly identify misalignment.
- Potential extension's toward population-wide moral shifts

Thank You!

Questions?