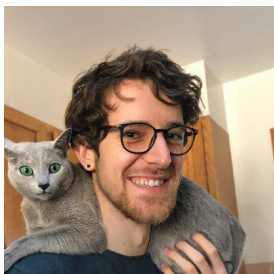


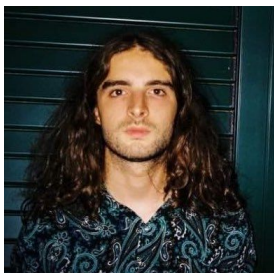
Beyond Preferences in AI Alignment

Towards Richer Models of
Human Reasons & Decisions

Xuan (Tan Zhi Xuan), *MIT*



Micah Carroll



Matija Franklin



Hal Ashton

<https://arxiv.org/abs/2408.16984>

Beyond Preferences in AI Alignment

Tan Zhi-Xuan
MIT

Micah Carroll
UC Berkeley

Matija Franklin
University College London

Hal Ashton
University of Cambridge

Abstract

The dominant practice of AI alignment assumes (1) that preferences are an adequate representation of human values, (2) that human rationality can be understood in terms of maximizing the satisfaction of preferences, and (3) that AI systems should be aligned with the preferences of one or more humans to ensure that they behave safely and in accordance with our values. Whether implicitly followed or explicitly endorsed, these commitments constitute what we term a *preferentist* approach to AI alignment. In this paper, we characterize and challenge the preferentist approach, describing conceptual and technical alternatives that are ripe for further research. We first survey the limits of rational choice theory as a descriptive model, explaining how preferences fail to capture the thick semantic content of human values, and how utility representations neglect the possible incommensurability of those values. We then critique the normativity of expected utility theory (EUT) for humans and AI, drawing upon arguments showing how rational agents need not comply with EUT, while highlighting how EUT is silent on which preferences are normatively acceptable. Finally, we argue that these limitations motivate a reframing of the targets of AI alignment: Instead of alignment with the preferences of a human user, developer, or humanity-writ-large, AI systems should be aligned with normative standards appropriate to their social roles, such as the role of a general-purpose assistant. Furthermore, these standards should be negotiated and agreed upon by all relevant stakeholders. On this alternative conception of alignment, a multiplicity of AI systems will be able to serve diverse ends, aligned with normative standards that promote mutual benefit and limit harm despite our plural and divergent values.

Zhi-Xuan et al (in press), *Philosophical Studies*, Special Issue on AI Safety.

AI Alignment vs. AI Safety

AI alignment: The project of ensuring that intelligent autonomous systems robustly act in our (collective) interests.

AI safety: The project of ensuring that intelligent autonomous systems avoid (catastrophic) loss or harm to people or society.

AI Alignment vs. AI Safety

AI alignment: The project of ensuring that intelligent autonomous systems robustly act in our (collective) interests.



Since our interests include avoiding (catastrophic) harm, safety is a minimal requirement for alignment.

AI safety: The project of ensuring that intelligent autonomous systems avoid (catastrophic) loss or harm to people or society.

AI Alignment vs. AI Safety

AI alignment: The project of ensuring that intelligent autonomous systems robustly act in our (collective) interests.



If “AI” = “powerful expected utility maximizer”, then safety requires alignment (i.e. maximizing the right thing).

AI safety: The project of ensuring that intelligent autonomous systems avoid (catastrophic) loss or harm to people or society.

AI Alignment vs. AI Safety

If “AI” = “powerful expected utility maximizer”, then safety requires alignment (i.e. maximizing the right thing).

AI Alignment vs. AI Safety

If “AI” = “powerful expected utility maximizer”, then safety requires alignment (i.e. maximizing the right thing).



*Since AI has to maximize a utility function to be safe, it should maximize the **human** utility function.*



Utility functions are just preference orderings over outcomes that adhere to certain postulates of rationality.

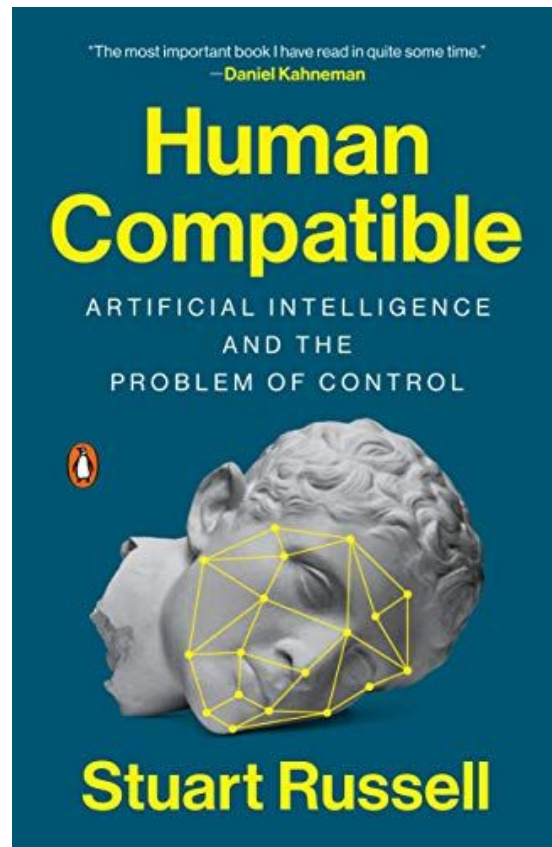


For AI systems to be safe, they should be aligned so as to maximize the satisfaction of human preferences.

Preferentism in AI Alignment

Russell's Principles for Beneficial AI:

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.



Preferentism in AI Alignment

Learning What to Value

Daniel Dewey

Machine Intelligence Research Institute

Abstract. We examine ultraintelligent reinforcement learning agents. Reinforcement learning can only be used in the real world to define agents whose goal is to maximize expected rewards, and since this goal does not match with human goals, AGIs based on reinforcement learning will often work at cross-purposes to us. We define value learners, agents that can be designed to learn and maximize any initially unknown utility function so long as we provide them with an idea of what constitutes evidence about that utility function.

(Dewey, 2011)

The AI Alignment Problem: Why It's Hard, and Where to Start

Eliezer Yudkowsky
Machine Intelligence Research Institute
eliezer@intelligence.org

1 Agents and their utility functions

In this talk, I'm going to try to answer the frequently asked question, "Just what is it that you do all day long?" As a starting frame, I'd like to say that before you try to persuade anyone of something, you should first try to make sure that they know what the heck you're talking about. It is in that spirit that I'd like to offer this talk. Persuasion can come during Q&A. If you have a disagreement, hopefully I can address it during Q&A. The purpose of this talk is to have you understand what this field is about, so that you can disagree with it.

First, "The primary concern," said Stuart Russell, "is not not spooky emergent consciousness but simply the ability to make *high-quality decisions*." We are concerned with the theory of artificial intelligences that are advanced beyond the present day, and that make sufficiently high-quality decisions in the service of whatever goals (or, in particular, utility functions) they may have been programmed with to be objects of concern.

Coherent decisions imply a utility function

(Yudkowsky, 2016)

Preferentism in AI Alignment

Cooperative Inverse Reinforcement Learning

Dylan Hadfield-Menell* Anca Dragan Pieter Abbeel Stuart Russell
Electrical Engineering and Computer Science
University of California at Berkeley
Berkeley, CA 94709

Abstract

For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans. We propose a formal definition of the value alignment problem as *cooperative inverse reinforcement learning* (CIRL). A CIRL problem is a cooperative, partial-information game with two agents, human and robot; both are rewarded according to the human's reward function, but the robot does not initially know what this is. In contrast to classical IRL, where the human is assumed to act optimally in isolation, optimal CIRL solutions produce behaviors such as active teaching, active learning, and communicative actions that are more effective in achieving value alignment. We show that computing optimal joint policies in CIRL games can be reduced to solving a POMDP, prove that optimality in isolation is suboptimal in CIRL, and derive an approximate CIRL algorithm.

(Hadfield-Menell et al, 2016)

Deep Reinforcement Learning from Human Preferences

Paul F Christiano Jan Leike Tom B Brown
OpenAI DeepMind Google Brain*
paul@openai.com leike@google.com tombrown@google.com

Miljan Martic Shane Legg Dario Amodei
DeepMind DeepMind OpenAI
miljanm@google.com legg@google.com damodei@openai.com

Abstract

For sophisticated reinforcement learning (RL) systems to interact usefully with real-world environments, we need to communicate complex goals to these systems. In this work, we explore goals defined in terms of (non-expert) human preferences between pairs of trajectory segments. We show that this approach can effectively solve complex RL tasks without access to the reward function, including Atari games and simulated robot locomotion, while providing feedback on less than 1% of our agent's interactions with the environment. This reduces the cost of human oversight far enough that it can be practically applied to state-of-the-art RL systems. To demonstrate the flexibility of our approach, we show that we can successfully train complex novel behaviors with about an hour of human time. These behaviors and environments are considerably more complex than any which have been previously learned from human feedback.

(Christiano et al, 2017)

Preferentism in AI Alignment

Training language models to follow instructions with human feedback

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through a language model API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

(OpenAI, 2022)

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov^{*†}

Archit Sharma^{*†}

Eric Mitchell^{*†}

Stefano Ermon^{†‡}

Christopher D. Manning[†]

Chelsea Finn[†]

[†]Stanford University [‡]CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Abstract

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper, we leverage a mapping between reward functions and optimal policies to show that this constrained reward maximization problem can be *optimized exactly* with a single stage of policy training, essentially solving a classification problem on the human preference data. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for fitting a reward model, sampling from the LM during fine-tuning, or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds RLHF's ability to control sentiment of generations and improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

(Rafailov, Sharma & Mitchell et al, 2023)

Preferentism in AI Alignment

An approach to AI alignment that treats preferences (or “reward”, or “utility”) as:

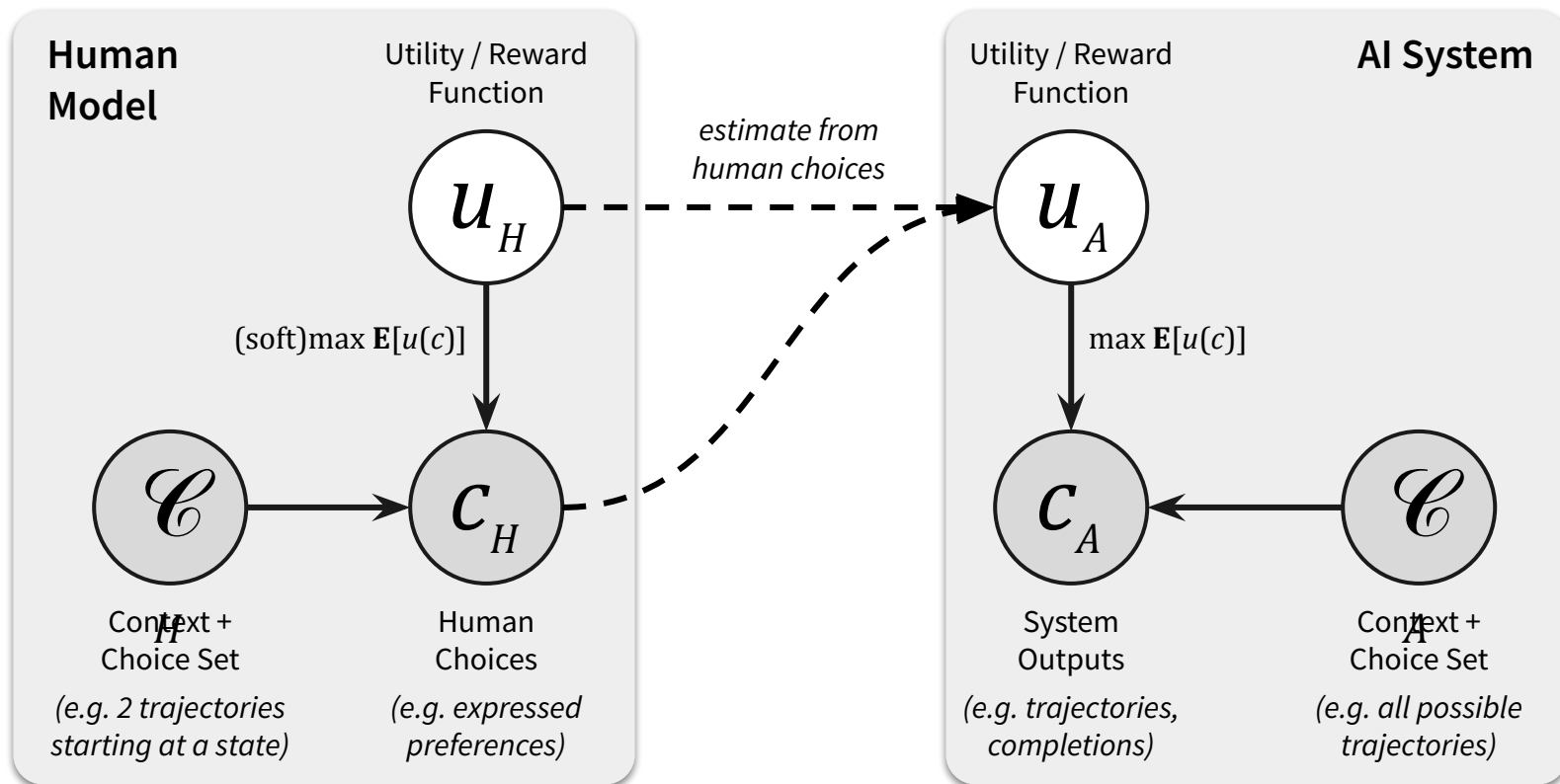
- **Ontologically adequate:** Preferences/reward/utility fully define the content of human values, task specifications, or value-aligned behavior.
- **Epistemically central:** Preferences/reward/utility are what AI systems need to learn in order to understand and produce aligned behavior.
- **Normatively basic:** Satisfying human preferences or maximizing human utility is the ultimate normative standard for judging whether an AI system is aligned.

Preferentism in AI Alignment

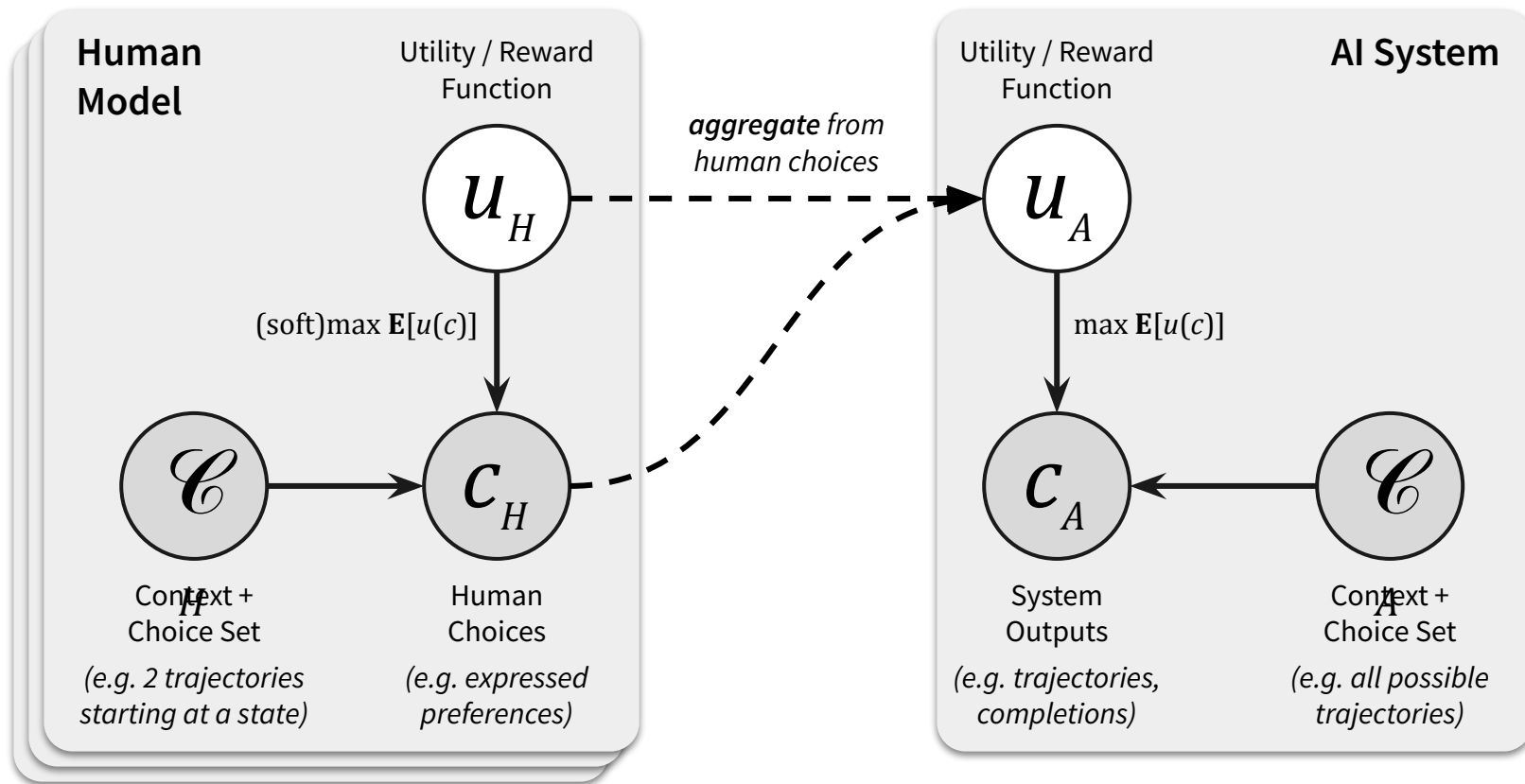
Dominant methods, frameworks, and formalizations of AI (mis)alignment typically assume one or more of the following theses:

- **Rational Choice Theory as a Descriptive Account.** Human decisions are well-modeled as approx. maximizing the satisfaction of preferences, which can be represented as a utility or reward function.
- **Expected Utility Theory as a Normative Standard.** Rationality can be characterized as the maximization of expected utility, and AI should be designed & analyzed according to this standard.
- **Single-Agent Alignment as Preference Matching.** For an AI system to be aligned to a single human, it should act so as to maximize the satisfaction of the preferences of that human.
- **Multi-Agent Alignment as Preference Aggregation.** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

The Preferentist Model of Humans and AI



The Preferentist Model of Humans and AI



Beyond Preferentism in AI Alignment

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- **Rational Choice Theory as a Descriptive Account.** Human decisions are well-modeled as approx. maximizing the satisfaction of preferences, which can be represented as a utility or reward function.
- **Expected Utility Theory as a Normative Standard.** Rationality can be characterized as the maximization of expected utility, and AI should be designed & analyzed according to this standard.
- **Single-Agent Alignment as Preference Matching.** For an AI system to be aligned to a single human, it should act so as to maximize the satisfaction of the preferences of that human.
- **Multi-Agent Alignment as Preference Aggregation.** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

Beyond Preferentism in AI Alignment

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- **Expected Utility Theory as a Normative Standard.** Rationality can be characterized as the maximization of expected utility, and AI should be designed & analyzed according to this standard.
- **Single-Agent Alignment as Preference Matching.** For an AI system to be aligned to a single human, it should act so as to maximize the satisfaction of the preferences of that human.
- **Multi-Agent Alignment as Preference Aggregation.** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

***Beyond* Preferentism in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory.*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Single-Agent Alignment as Preference Matching.*** For an AI system to be aligned to a single human, it should act so as to maximize the satisfaction of the preferences of that human.
- ***Multi-Agent Alignment as Preference Aggregation.*** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

***Beyond* Preferentism in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory:*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Beyond Single-Agent Alignment as Preference Matching:*** Alignment with *task or role-specific normative criteria*, such as *the normative ideal for a (general-purpose AI) assistant*.
- ***Multi-Agent Alignment as Preference Aggregation.*** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

***Beyond* Preferentism in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

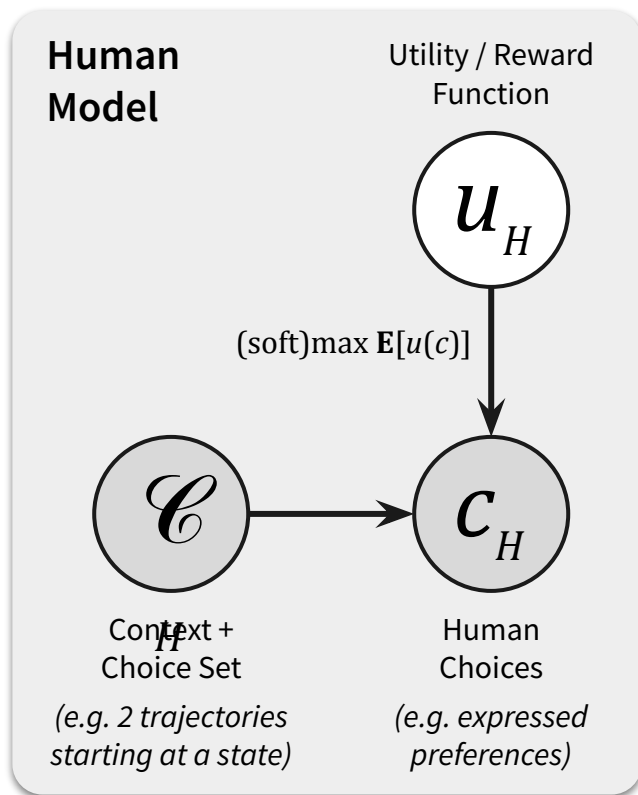
- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory:*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Beyond Single-Agent Alignment as Preference Matching:*** Alignment with *task or role-specific normative criteria*, such as *the normative ideal for a (general-purpose AI) assistant*.
- ***Beyond Multi-Agent Alignment as Preference Aggregation:*** Alignment with *a plurality of normative standards* for a *plurality of AI systems*, given our *plural and divergent interests*.

***Beyond* Preferentism in AI Alignment**

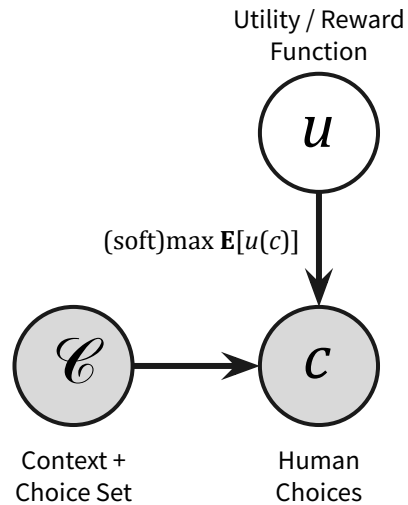
We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

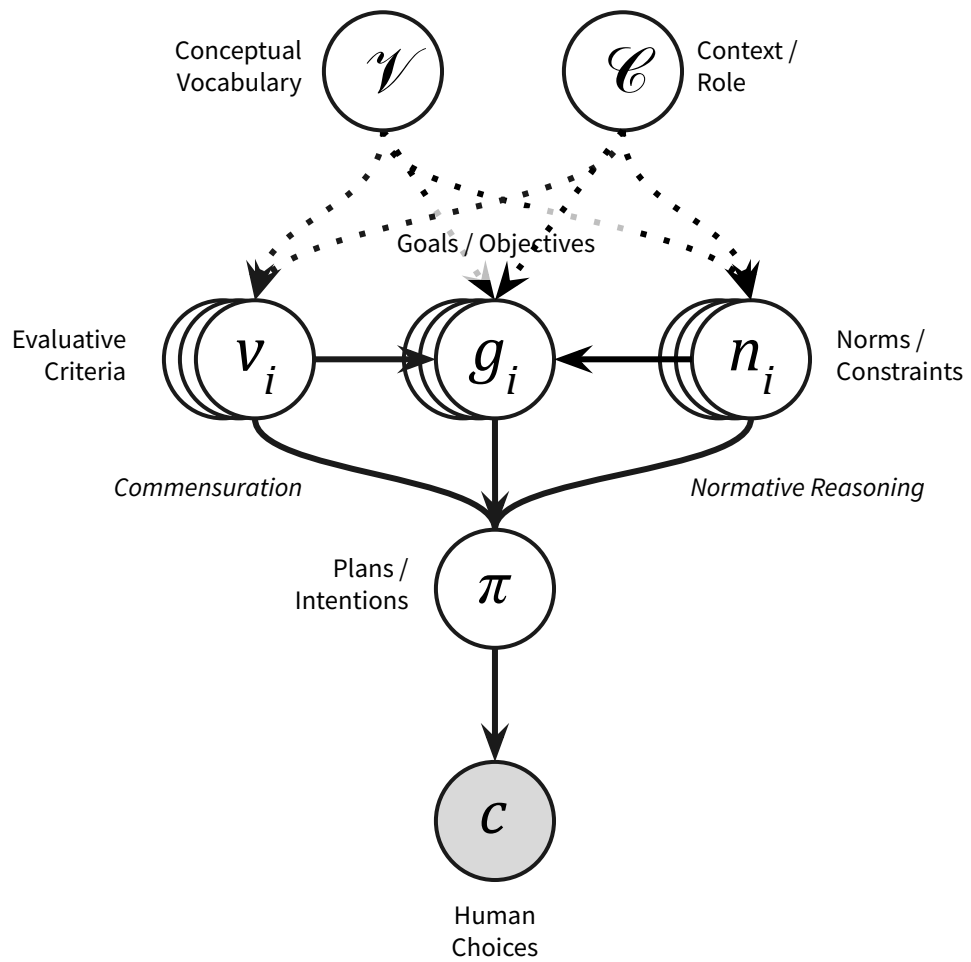
- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory:*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses, design targets*, and *richer theories of (human) reason*.
- ***Beyond Single-Agent Alignment as Preference Matching:*** Alignment with *task or role-specific normative criteria*, such as *the normative ideal for a (general-purpose AI) assistant*.
- ***Beyond Multi-Agent Alignment as Preference Aggregation:*** Alignment with *a plurality of normative standards* for a *plurality of AI systems*, given our *plural and divergent interests*.

Beyond the Preferentist Model



Beyond the Preferentist Model





Beyond Preferentism in AI Alignment

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- **Expected Utility Theory as a Normative Standard.** Rationality can be characterized as the maximization of expected utility, and AI should be designed & analyzed according to this standard.
- **Single-Agent Alignment as Preference Matching.** For an AI system to be aligned to a single human, it should act so as to maximize the satisfaction of the preferences of that human.
- **Multi-Agent Alignment as Preference Aggregation.** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

Rational Choice Theory as a *Descriptive Account*

- Assumes human preferences can be represented by a utility function, and that humans make choices by maximizing expected utility:

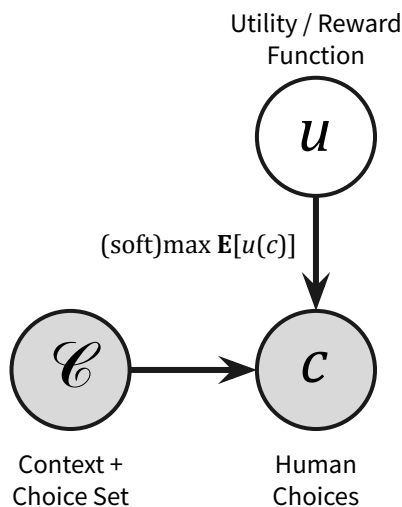
$$c^* = \operatorname{argmax}_c \mathbf{E}[U(c)]$$

- In AI & machine learning, strict optimality is often relaxed, giving *noisy/Boltzmann rationality**:

$$P(c) \propto \exp(\mathbf{E}[U(c)])$$

- Can be extended to all forms of human feedback via the framework of *reward-rational implicit choice* (Jeon, Milli & Dragan, 2020).

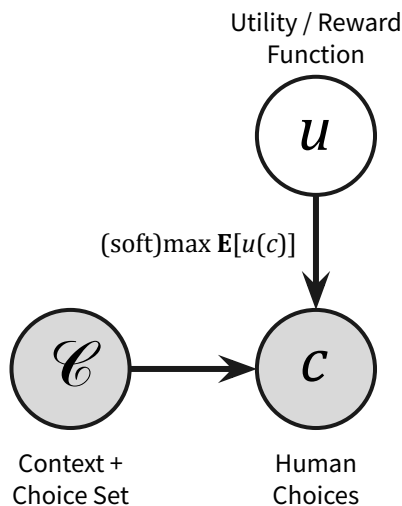
*a.k.a. Plackett-Luce, Max-Entropy RL, Random Utility Model w Gumbel Noise



Beyond Rational Choice Theory

Beyond noisily-rational models of human decisions

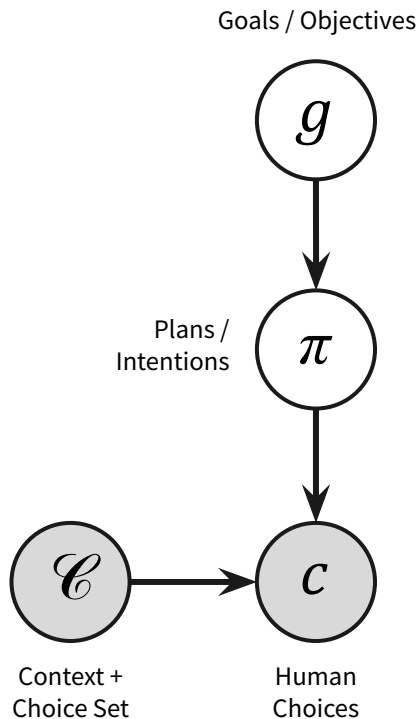
- Humans are not just noisily-rational, but *boundedly rational*, subject to cognitive biases and limitations.
- *In RLHF*, people may use response length as a heuristic for helpfulness, or prefer intuitive but subtly incorrect reasoning.
- *In inverse RL*, people may provide sub-optimal demonstrations for hard planning problems (e.g. chess, traveling salesperson, Sokoban).



Beyond Rational Choice Theory

Beyond noisily-rational models of human decisions

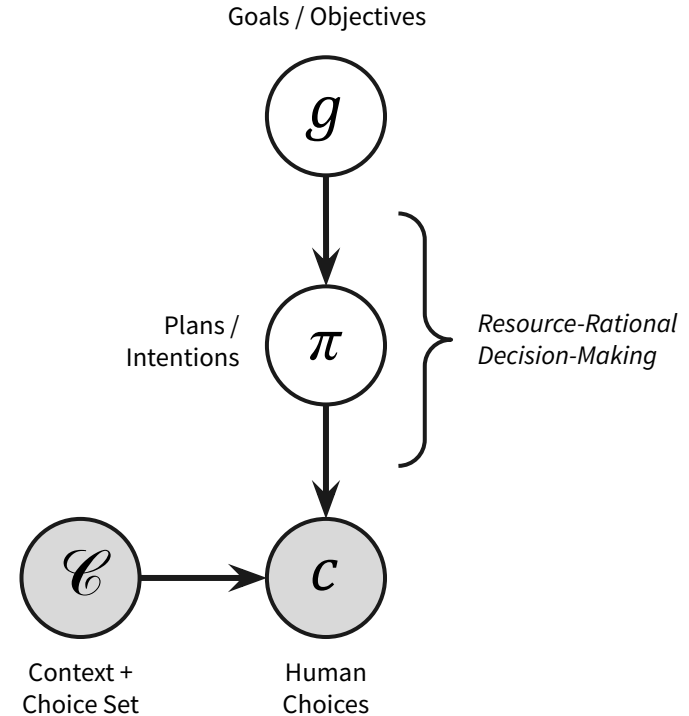
- Humans are not just noisily-rational, but *boundedly rational*, subject to cognitive biases and limitations.
- *In RLHF*, people may use response length as a heuristic for helpfulness, or prefer intuitive but subtly incorrect reasoning.
- *In inverse RL*, people may provide sub-optimal demonstrations for hard planning problems (e.g. chess, traveling salesperson, Sokoban).



Beyond Rational Choice Theory

Beyond noisily-rational models of human decisions

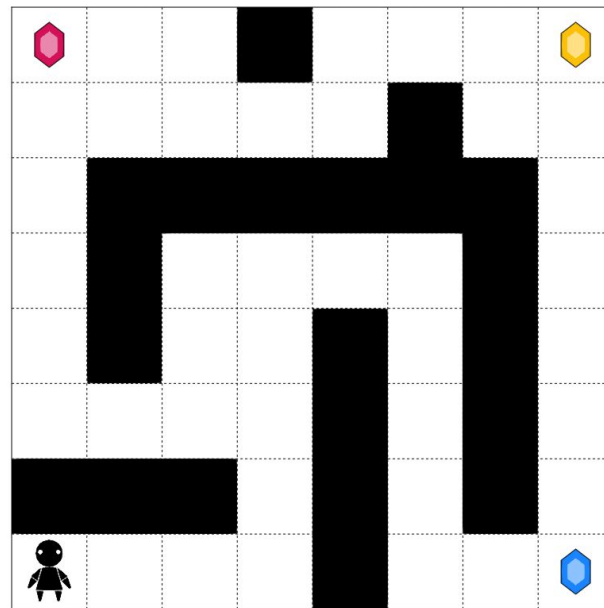
- *Resource rationality* — the rational use of limited cognitive resources (Lieder & Griffiths, 2020) — can guide the design of better human models.
- Tries to make sense of humans as *rational creatures, but forgivingly*.
- Resource-rational planning: Model humans as *thinking ahead* before acting, but only for a *limited amount of steps*.



Beyond Rational Choice Theory

Beyond noisily-rational models of human decisions

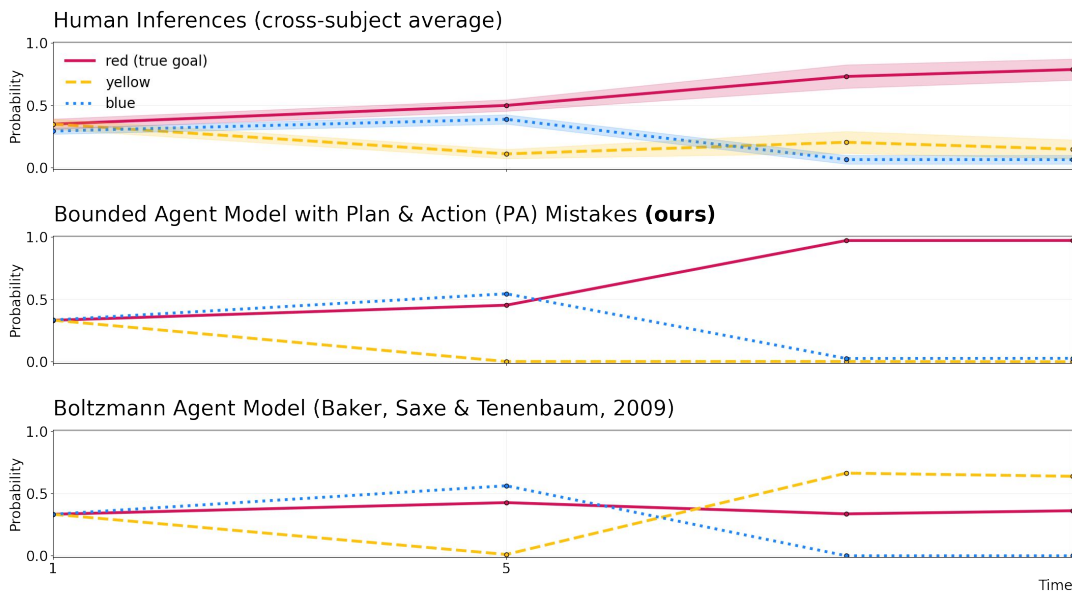
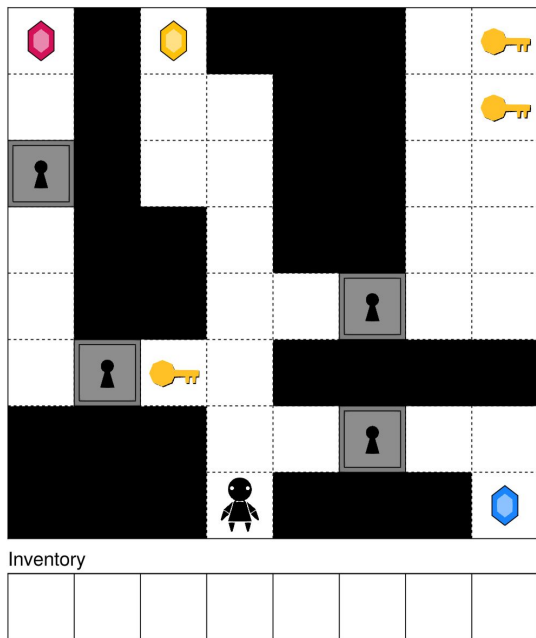
- *Resource rationality* — the rational use of limited cognitive resources — can guide us towards the design of better human models.
- Tries to make sense of humans as *rational creatures, but forgivingly*.
- Resource-rational planning: Model humans as *thinking ahead* before acting, but only for a *limited amount of steps*.



(Zhi-Xuan et al, NeurIPS 2020)

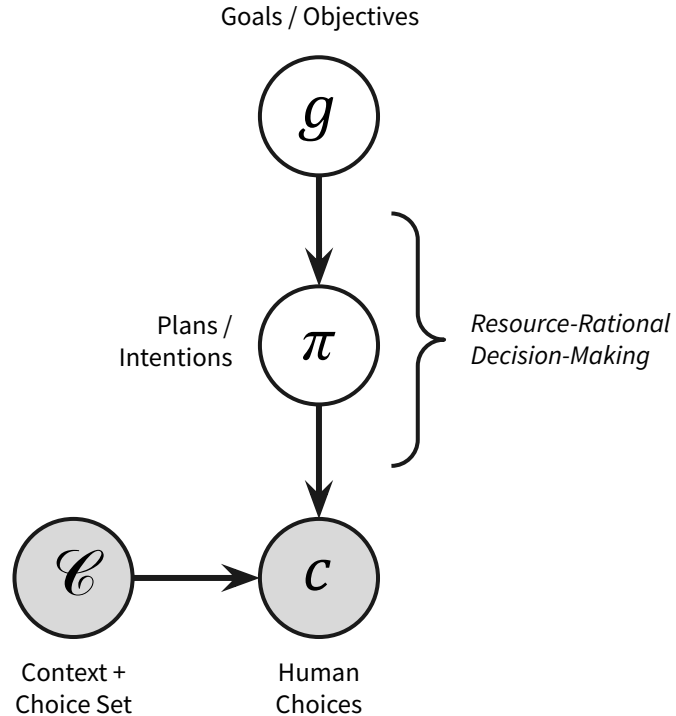
Beyond Rational Choice Theory

Beyond noisily-rational models of human decisions



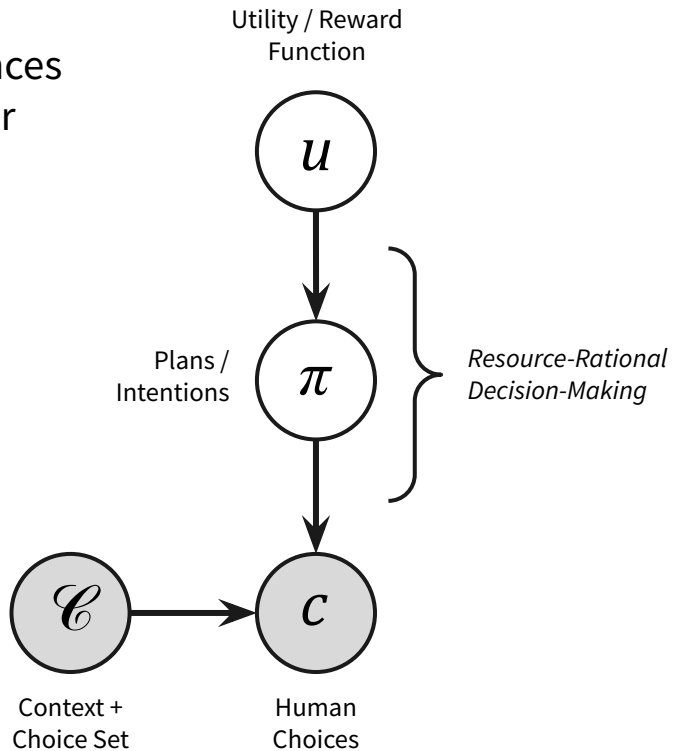
(Alanqary, Lin, Le, Zhi-Xuan et al, CogSci 2021)

Beyond Rational Choice Theory



Beyond Rational Choice Theory

Can't human goals or preferences still be represented as utility or reward functions?



Beyond Rational Choice Theory

Beyond reward and utility representations

Reward is enough

David Silver*, Satinder Singh, Doina Precup, Richard S. Sutton



ARTICLE INFO

Article history:

Received 12 November 2020

Received in revised form 28 April 2021

Accepted 12 May 2021

Available online 24 May 2021

Keywords:

Artificial intelligence

Artificial general intelligence

Reinforcement learning

Reward

ABSTRACT

In this article we hypothesise that intelligence, and its associated abilities, can be understood as subserving the maximisation of reward. Accordingly, reward is enough to drive behaviour that exhibits abilities studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language, generalisation and imitation. This is in contrast to the view that specialised problem formulations are needed for each ability, based on other signals or objectives. Furthermore, we suggest that agents that learn through trial and error experience to maximise reward could learn behaviour that exhibits most if not all of these abilities, and therefore that powerful reinforcement learning agents could constitute a solution to artificial general intelligence.

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Beyond Rational Choice Theory

Beyond reward and utility representations

- **Opacity:** Utility/reward functions obscure the *underlying semantics* of human goals, values, and reasons, which are defined in a rich conceptual language.
- **Scalarity:** Utility/reward functions assume that we can *always* commensurate our goals and values into a single scalar value (i.e. that our preferences are *complete*).

Beyond Rational Choice Theory

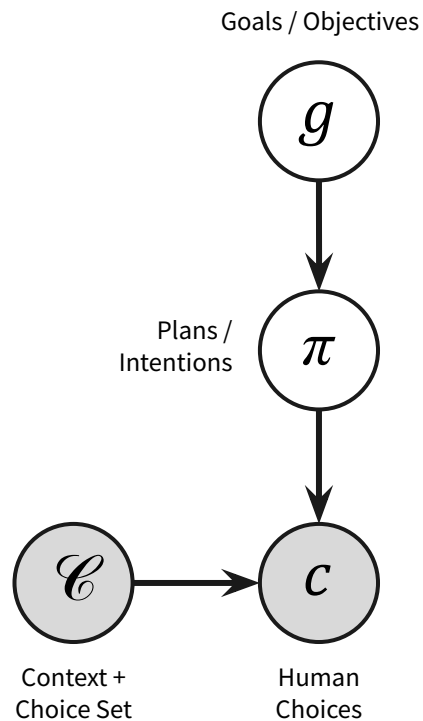
Beyond reward and utility representations

- **Opacity:** Utility/reward functions obscure the *underlying semantics* of human goals, values, and reasons, which are defined in a rich conceptual language.
- **Scalarity:** Utility/reward functions assume that we can *always* commensurate our goals and values into a single scalar value (i.e. that our preferences are *complete*).

Beyond Rational Choice Theory

Beyond reward and utility representations

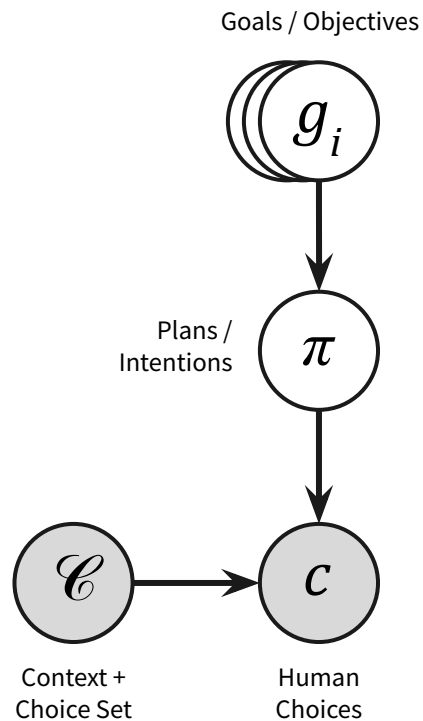
- I want to get a paper into NeurIPS



Beyond Rational Choice Theory

Beyond reward and utility representations

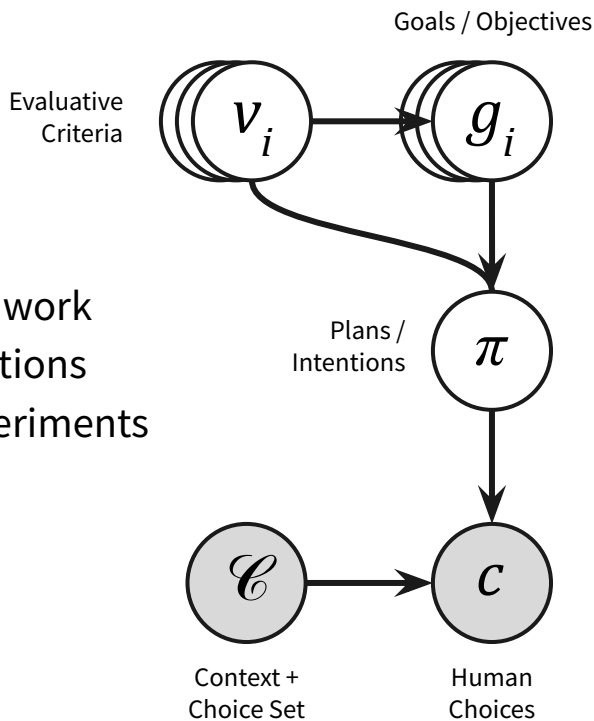
- I want to get a paper into NeurIPS
 - But also not lose too much sleep
 - And also spend time with friends



Beyond Rational Choice Theory

Beyond reward and utility representations

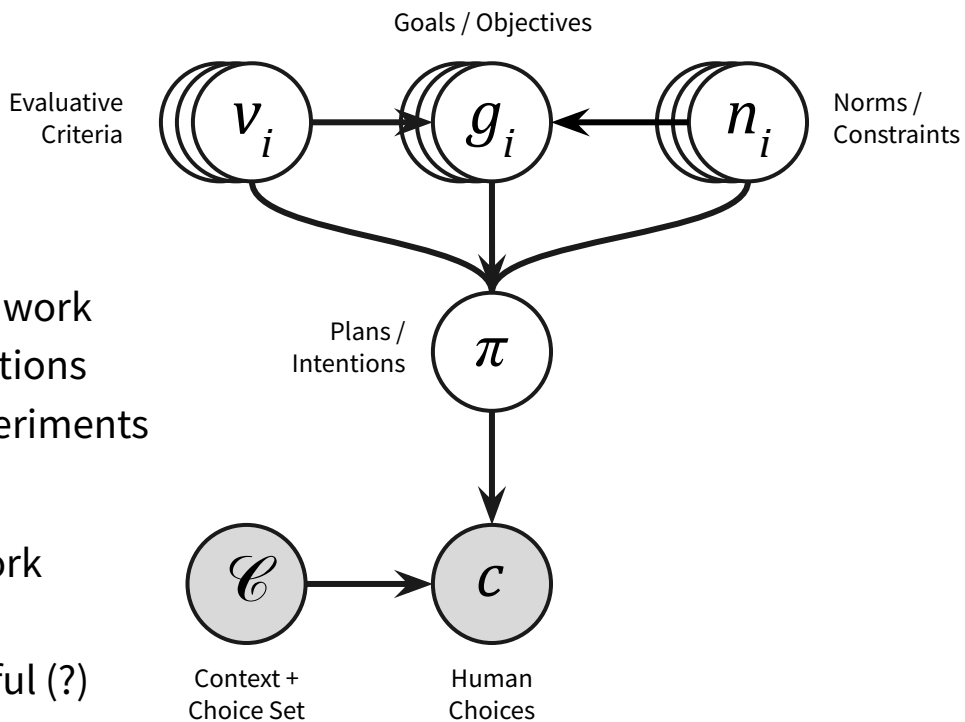
- I want to get a paper into NeurIPS...
 - But also not lose too much sleep
 - And also spend time with friends
- The paper should be:
 - *Novel* relative to the existing field of work
 - *Impactful* in its downstream implications
 - *Technically sound* in theory and experiments



Beyond Rational Choice Theory

Beyond reward and utility representations

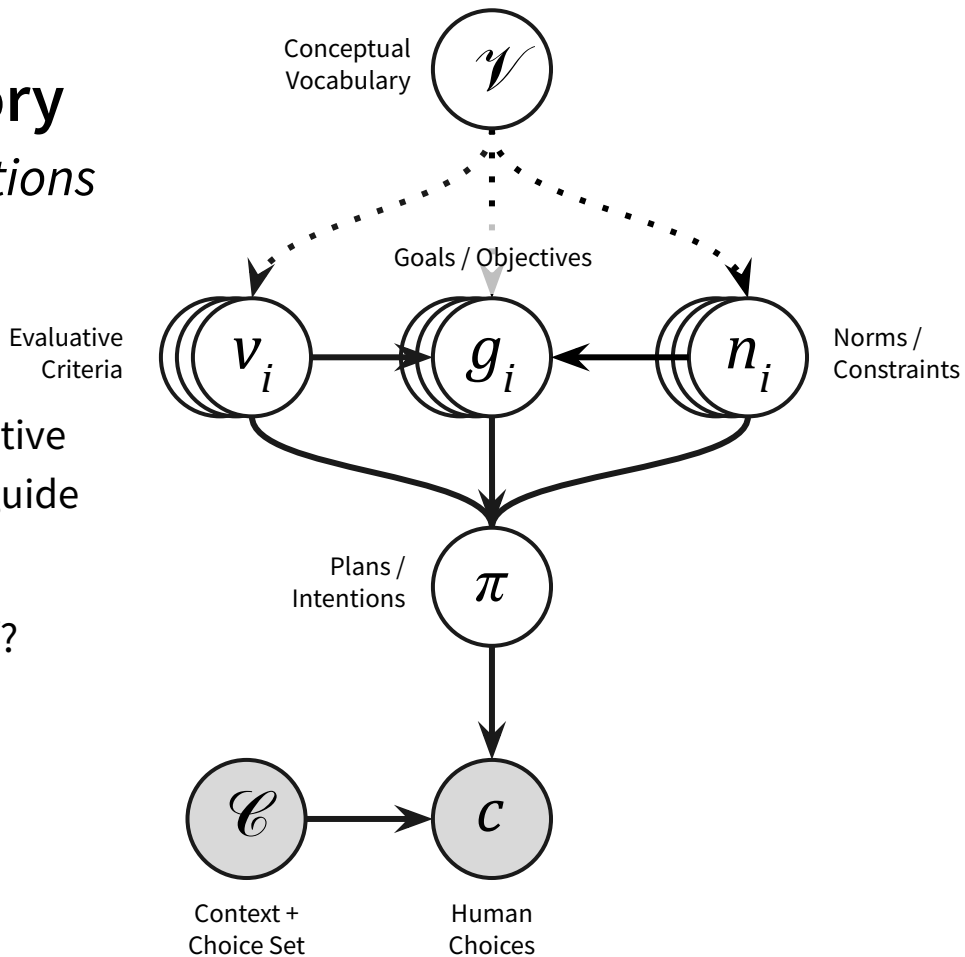
- I want to get a paper into NeurIPS...
 - But also not lose too much sleep
 - And also spend time with friends
- The paper should be:
 - *Novel* relative to the existing field of work
 - *Impactful* in its downstream implications
 - *Technically sound* in theory and experiments
- I should:
 - Avoid plagiarizing other people's work
 - Ensure my work is reproducible
 - Avoid research that is socially harmful (?)



Beyond Rational Choice Theory

Beyond reward and utility representations

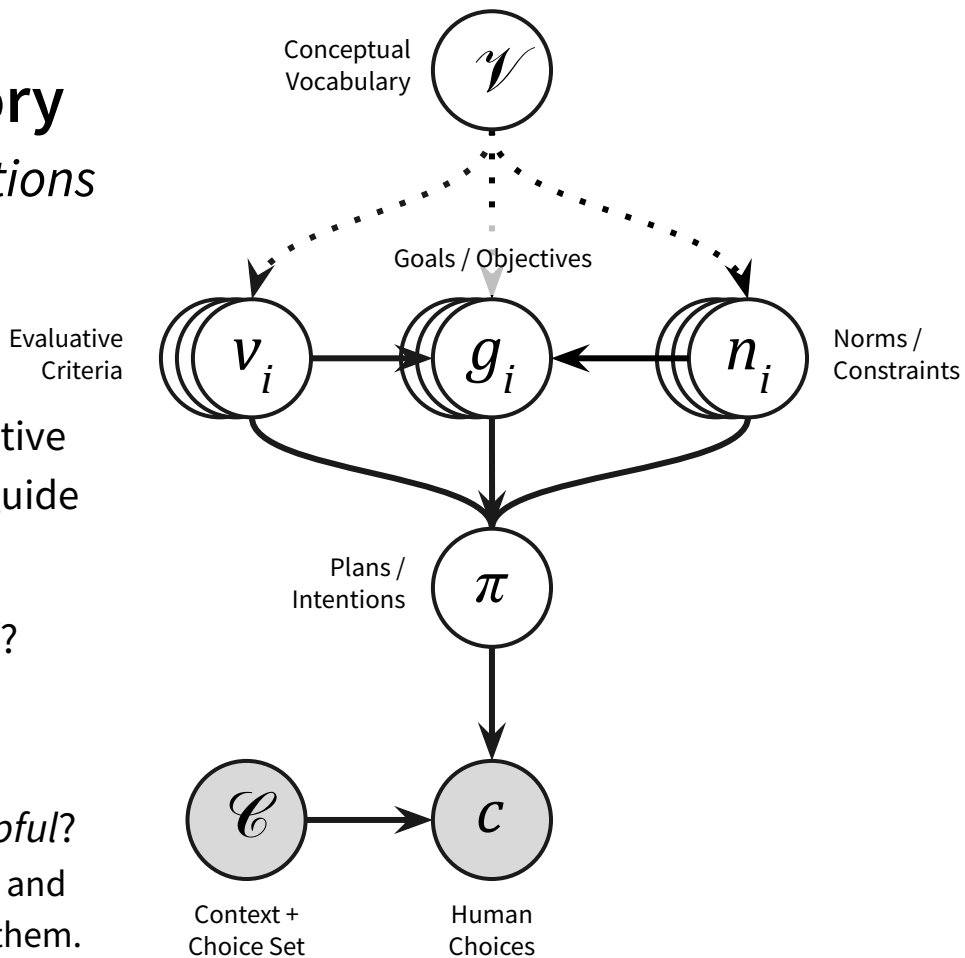
- Each criterion or constraint corresponds to an *evaluative / normative concept*.
- Humans learn a rich vocabulary of evaluative concepts (i.e. values) and apply them to guide action or judge aspects of the world.
- What does it mean for a paper to be *novel*?
 - Requires surveying the existing field and knowing the gaps in the literature.



Beyond Rational Choice Theory

Beyond reward and utility representations

- Each criterion or constraint corresponds to an *evaluative / normative concept*.
- Humans learn a rich vocabulary of evaluative concepts (i.e. values) and apply them to guide action or judge aspects of the world.
- What does it mean for a paper to be *novel*?
 - Requires surveying the existing field and knowing the gaps in the literature.
- What does it mean for an action to be *helpful*?
 - Requires figuring out other agent's goals, and checking if the action enables achieving them.



Beyond Rational Choice Theory

Beyond reward and utility representations

- Each criterion or constraint corresponds to an *evaluative / normative concept*.
- Humans learn a rich vocabulary of evaluative concepts (i.e. values) and apply them to guide action or judge aspects of the world.
- What does it mean for a paper to be *novel*?
 - Requires surveying the existing field and knowing the gaps in the literature.
- What does it mean for an action to be *helpful*?
 - Requires figuring out other agent's goals, and checking if the action enables achieving them.

Value as Semantics: Representations of Human Moral and Hedonic Value in Large Language Models

Anna Leshinskaya*
AI Objectives Institute
San Francisco, CA
anna.leshinskaya@gmail.com

Aleksandr Chakroff
AI Objectives Institute
San Francisco, CA
alekchakroff@gmail.com

Abstract

Aligning AI with human objectives can be facilitated by enabling it to learn and veridically represent our values. In modern AI agents, value is a scalar magnitude reflecting the desirability of a given state or action. We propose a framework, value-as-semantics, in which such magnitudes are represented within a large-scale, high-dimensional semantic representation in a large language model. This approach allows value to be quantitative, yet also assigned to any expression in natural language and to inherit the expressivity and generalizability of the model's ontology. We used a broad set of action concepts to evaluate several assumptions of this approach. First, we showed that value representations could be retrieved from the language model distinctly from other attributes of the same actions and were closely correlated with that of human raters. We found that two psychologically distinct kinds of value, moral and hedonic, were also separable from each other to the same degree as in human raters, though we also found that moral and hedonic values were correlated in human ratings when using large sets of items. Finally, we showed that the value representations retrieved with our method reliably adapt to simple natural language evidence designed to elicit changes in values. Overall, we conclude that modern language models can effectively function as databases of human value. This value-as-semantics architecture can be an important contribution towards a broader, multi-faceted computational model of human-like action planning and moral reasoning.

Beyond Rational Choice Theory

Beyond reward and utility representations

- Each criterion or constraint corresponds to an *evaluative / normative concept*.
- Humans learn a rich vocabulary of evaluative concepts (i.e. values) and apply them to guide action or judge aspects of the world.
- What does it mean for a paper to be *novel*?
 - Requires surveying the existing field and knowing the gaps in the literature.
- What does it mean for an action to be *helpful*?
 - Requires figuring out other agent's goals, and checking if the action enables achieving them.

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

Anthropic

Abstract

We apply preference modeling and reinforcement learning from human feedback (RLHF) to finetune language models to act as helpful and harmless assistants. We find this alignment training improves performance on almost all NLP evaluations, and is fully compatible with training for specialized skills such as python coding and summarization. We explore an iterated online mode of training, where preference models and RL policies are updated on a weekly cadence with fresh human feedback data, efficiently improving our datasets and models. Finally, we investigate the robustness of RLHF training, and identify a roughly linear relation between the RL reward and the square root of the KL divergence between the policy and its initialization. Alongside our main results, we perform peripheral analyses on calibration, competing objectives, and the use of OOD detection, compare our models with human writers, and provide samples from our models using prompts appearing in recent related work.

(Anthropic, 2022)

Are LLMs or LLM-backed reward models really learning the semantics of these evaluative concepts? Or just some good enough approximation over the dataset?

Beyond Rational Choice Theory

Beyond reward and utility representations

- **Opacity:** Utility/reward functions obscure the *underlying semantics* of human goals, values, and reasons, which are defined in a rich conceptual language.
- **Scalarity:** Utility/reward functions assume that we can *always* commensurate our goals and values into a single scalar value (i.e. that our preferences are *complete*).

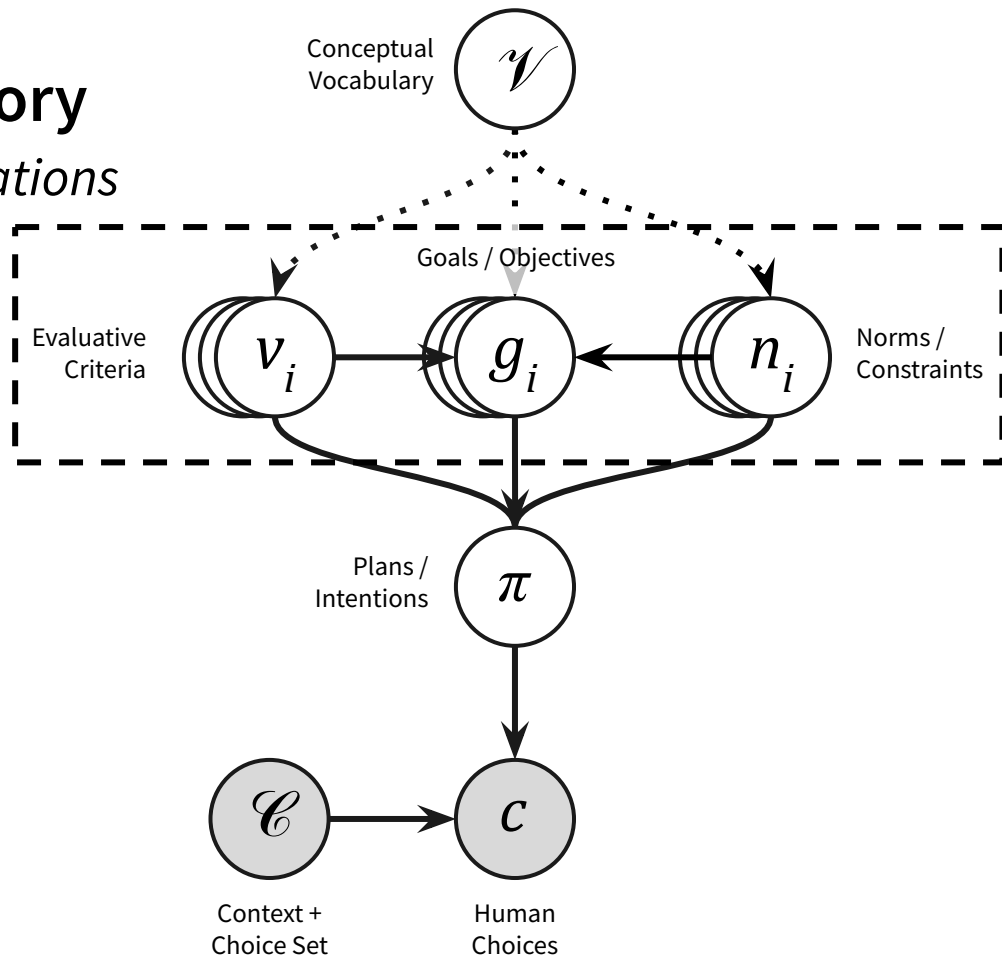
Beyond Rational Choice Theory

Beyond reward and utility representations

Can't I just compile this into a utility / reward function?

Especially if the reward function is defined over a space as rich as natural language?

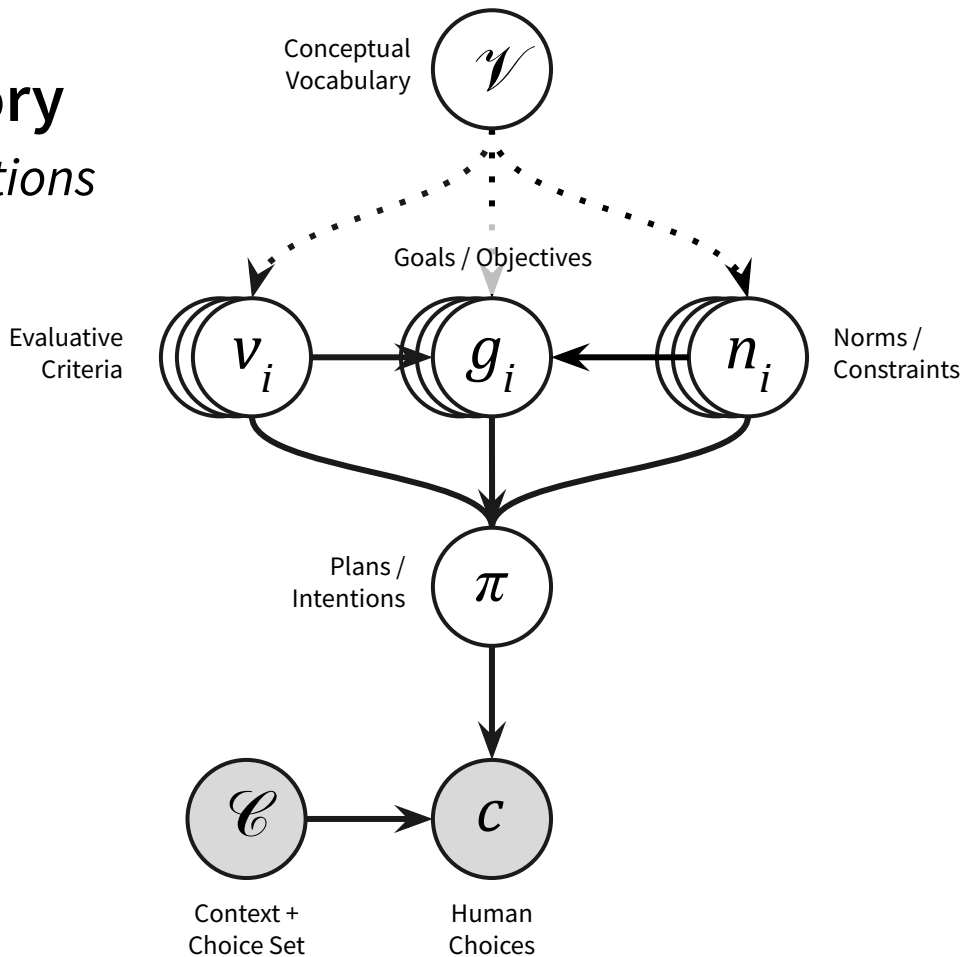
Especially if the reward function is just an LLM plus some scalar prediction head, so it captures all the semantics?



Beyond Rational Choice Theory

Beyond reward and utility representations

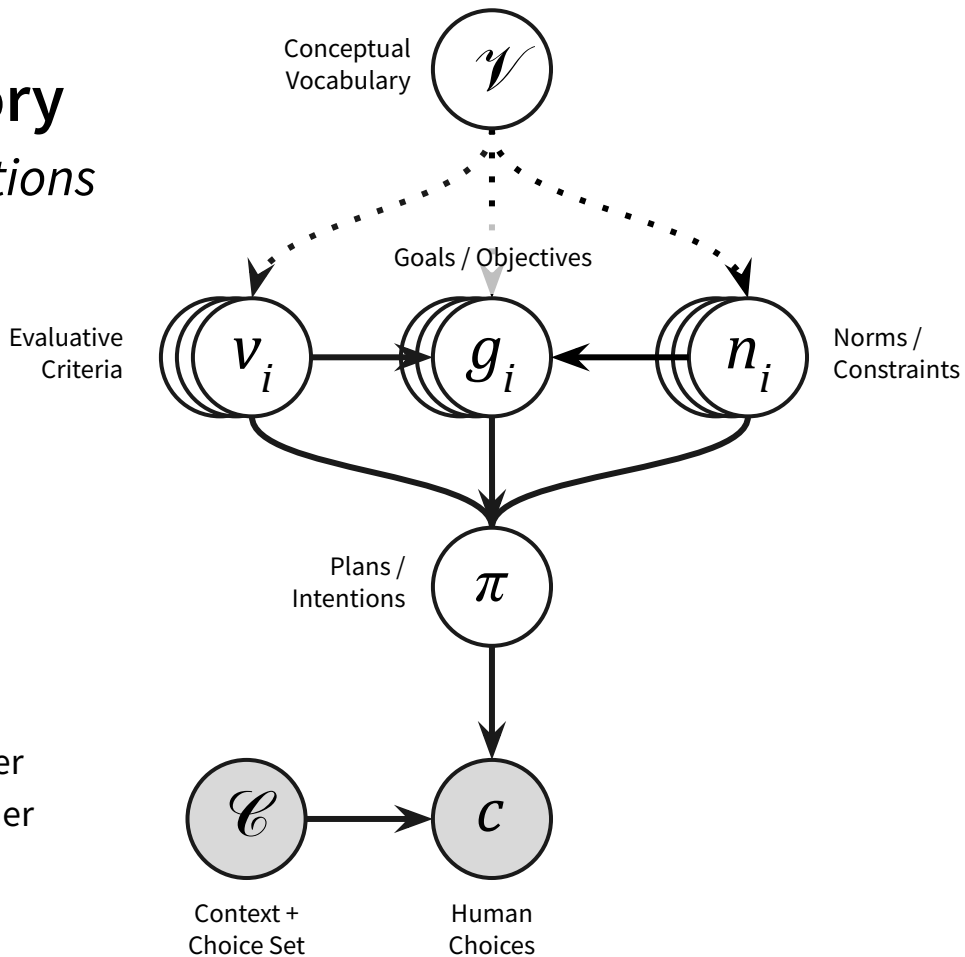
- **No.** Scalar utilities / rewards assume there are no *incomplete preferences*, due e.g. to *incommensurable values*.
- Some choices are *hard*! It doesn't seem like we can always say one choice is better than the other, or equally good.
- Which is better?
 - 1 NeurIPS paper + 1 sleepless night
 - 2 NeurIPS papers + 4 sleepless nights



Beyond Rational Choice Theory

Beyond reward and utility representations

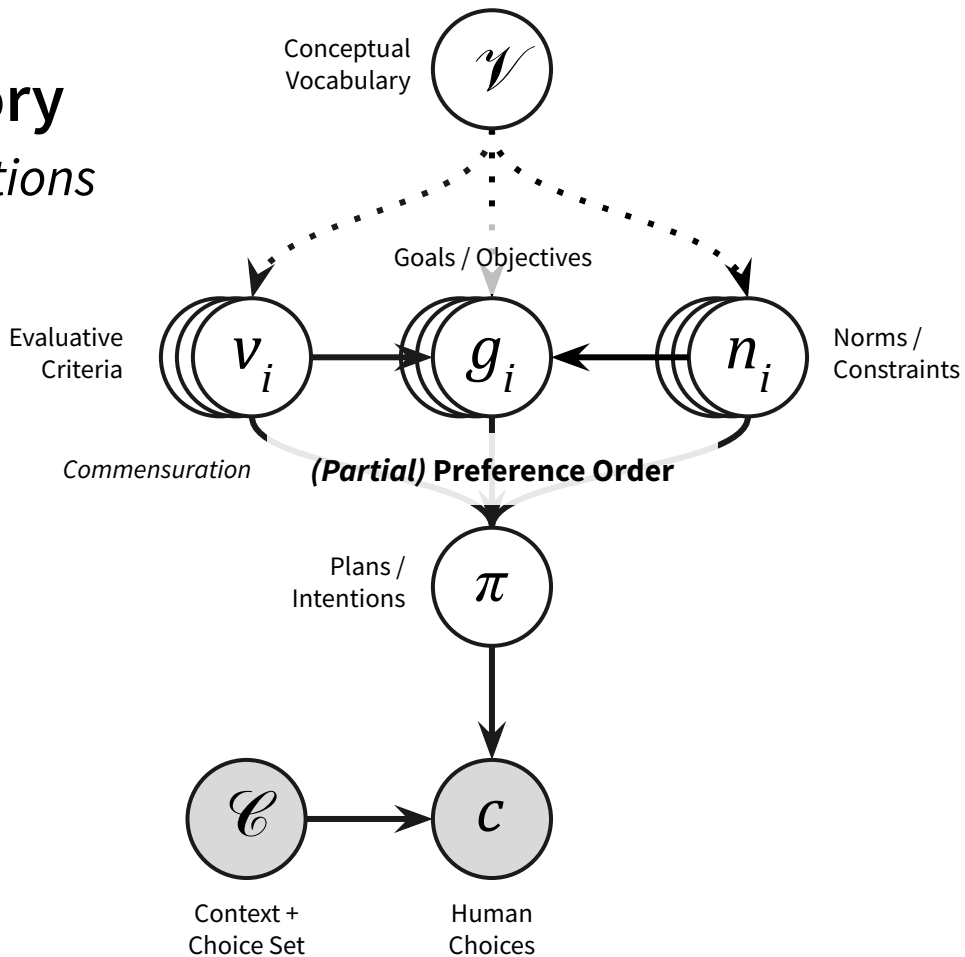
- **No.** Scalar utilities / rewards assume there are no *incomplete preferences*, due e.g. to *incommensurable values*.
- Some choices are *hard*! It doesn't seem like we can always say one choice is better than the other, or equally good.
- Which is better?
 - Job at dream school, far from your partner
 - Job at okay school, living with your partner



Beyond Rational Choice Theory

Beyond reward and utility representations

- **No.** Scalar utilities / rewards assume there are no *incomplete preferences*, due e.g. to *incommensurable values*.
- Some choices are *hard*! It doesn't seem like we can always say one choice is better than the other, or equally good.
- Which is better? Living in:
 - A wealthy country with no democracy
 - A poor country with democracy



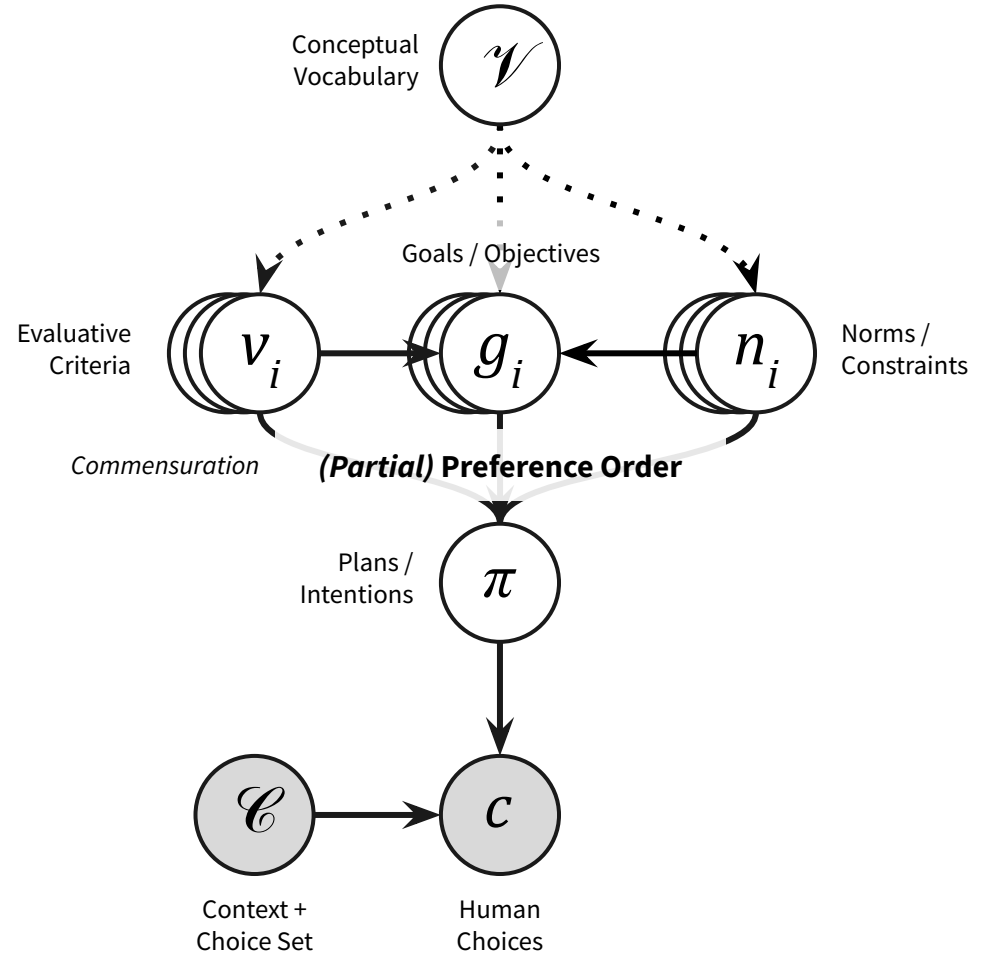
***Beyond* Preferentism in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- ***Beyond Rational Choice Theory***: Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory***. Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Single-Agent Alignment as Preference Matching***. For an AI system to be aligned to a single human, it should act so as to maximize the satisfaction of the preferences of that human.
- ***Multi-Agent Alignment as Preference Aggregation***. For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

Maybe humans can't be
described as expected utility
maximizers

But that's still what humans
and AI *should* try to do!



Expected Utility Theory as a Normative Standard

- A normative theory about what decisions are rational / how rational agents ought to act.
- Under certain axioms about what preferences count as rational, an agent with such preferences can be shown to act *as if* they are maximizing expected utility.
- In von Neumann & Morgenstern's (1944) representation theorem, these axioms are:
 - a. **Completeness:** For all outcome distributions A & B, either $A \geq B$ or $B \geq A$.
 - b. **Transitivity:** If $A \geq B$ and $B \geq C$, then $A \geq C$.
 - c. **Continuity:** If $A \geq B \geq C$, then there exists a distribution over A and C that's as good as B.
 - d. **Independence:** If $A \geq B$, then $A + pC \geq B + pC$ regardless of some extra alternative C.
- Similar axioms can be found in Savage's theory (1972), Bolker & Jeffrey's (1991), etc.

Expected Utility Theory as a Normative Standard

- In the AI alignment literature, such axioms are often taken as *requirements of rationality* that sufficiently advanced AI would adhere to.
- Typical justifications are *Dutch Book* or *money pump arguments*: Non-EU preferences are argued to be *exploitable* or *vulnerable*.

The AI Alignment Problem:
Why It's Hard, and Where to Start

Eliezer Yudkowsky
Machine Intelligence Research Institute
eliezer@intelligence.org

1 Agents and their utility functions

In this talk, I'm going to try to answer the frequently asked question, "Just what is it that you do all day long?" As a starting frame, I'd like to say that before you try to persuade anyone of something, you should first try to make sure that they know what the heck you're talking about. It is in that spirit that I'd like to offer this talk. Persuasion can come during Q&A. If you have a disagreement, hopefully I can address it during Q&A. The purpose of this talk is to have you understand what this field is about, so that you can disagree with it.

First, "The primary concern," said Stuart Russell, "is not not spooky emergent consciousness but simply the ability to make *high-quality decisions*." We are concerned with the theory of artificial intelligences that are advanced beyond the present day, and that make sufficiently high-quality decisions in the service of whatever goals (or, in particular, utility functions) they may have been programmed with to be objects of concern.

Coherent decisions imply a utility function

(Yudkowsky, 2016)

Beyond Expected Utility Theory

Beyond EUT as a requirement for sufficiently intelligent agents

- Agents with incomplete preferences can resist exploitation by money pumps (Thornley, 2023; Petersen; 2023)

Invulnerable Incomplete Preferences: A Formal Statement

by **Sami Petersen** 42 min read 30th Aug 2023 30 comments ...

Corrigibility MATS Program AI Frontpage

This article presents a few theorems about the invulnerability of agents with incomplete preferences. Elliott Thornley's (2023) proposed approach to the AI shutdown problem relies on these preferential gaps, but John Wentworth and David Lorell^o have argued that they make agents play strictly dominated strategies.^[1] I claim this is false.

Conclusion

With the right choice rule, we can guarantee the invulnerability—unexploitability and opportunism—of agents with incomplete preferences. I've proposed one such rule, Dynamic Strong Maximality, which nevertheless doesn't ask agents to pick against their preferences. What's more, the choice behaviour this rule induces is *not* representable as the agent having implicitly completed its preferences. Even under awareness growth, the extent to which the rule can effectively complete an agent's implied preferences is permanently bounded above. And with the framework provided, it's possible to make statements about which kinds of completions are possible, and in what cases.

(Petersen, 2023)

Beyond Expected Utility Theory

Beyond EUT as a requirement for sufficiently intelligent agents

- Agents with incomplete preferences can resist exploitation by money pumps (Thornley, 2023; Petersen; 2023)
- For “hard” utility functions, complying with EUT axioms is *intractable* (by reduction to MAX-2-SAT) (Camara, 2021)

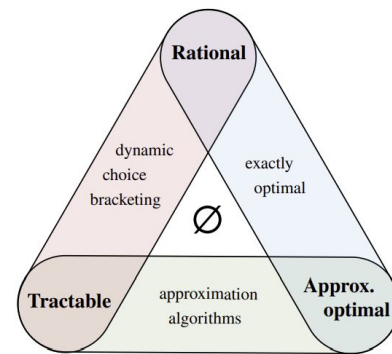


Figure 3: This diagram depicts the choice trilemma. The blue region connecting rationality and approximate optimality includes the traditional assumption of exact optimization. The green region connecting tractability and approximate optimality corresponds to approximation algorithms studied in computer science. The red region connecting rationality and tractability corresponds to dynamic choice bracketing. The \emptyset symbol says that the intersection of all three regions is empty.

Computationally Tractable Choice
(Camara, 2021)

Beyond Expected Utility Theory

Beyond EUT as a requirement for sufficiently intelligent agents

- Agents with incomplete preferences can resist exploitation by money pumps (Thornley, 2023; Petersen; 2023)
- For “hard” utility functions, complying with EUT axioms is *intractable* (by reduction to MAX-2-SAT) (Camara, 2021)
- Agents with non-EU preferences are *protected* by competitive markets (Laibson & Yariv, 2007), and can be *evolutionarily stable* (Widekind, 2008)

Safety in Markets: An Impossibility Theorem for Dutch Books*

David Laibson[†]
Harvard and NBER

Leeat Yariv[‡]
Caltech

Current Version: July 9, 2007

Abstract

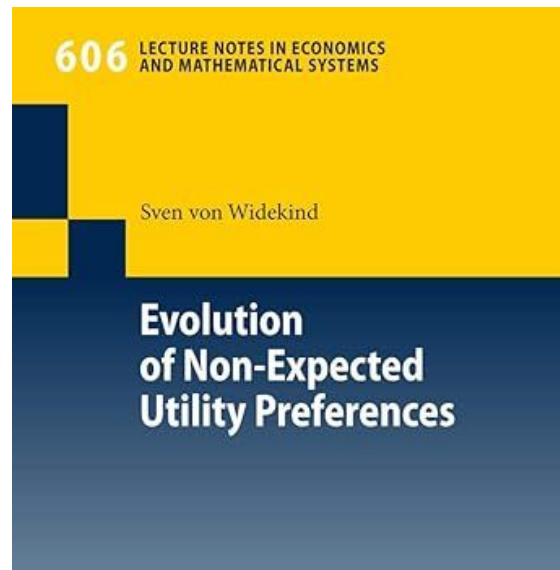
We show that competitive markets protect consumers from many forms of exploitation, even when consumers have non-standard preferences. We analyze a competitive dynamic economy in which consumers have arbitrary time-separable preferences and arbitrary beliefs about their own future behavior. Competition among agents eliminates rents and protects vulnerable consumers, who could have been exploited by a monopolist. In fact, in competitive general equilibrium no consumer participates in a trading sequence that strictly reduces her endowment – there are no Dutch Books. The absence of Dutch Books in and of itself does not distinguish standard and non-standard preferences. However, non-standard preferences do generate qualitatively different equilibrium outcomes than standard preferences. We characterize the testable implications of the standard model with a dynamic generalization of the Strong Axiom of Revealed Preferences.

(Laibson & Yariv, 2021)

Beyond Expected Utility Theory

Beyond EUT as a requirement for sufficiently intelligent agents

- Agents with incomplete preferences can resist exploitation by money pumps (Thornley, 2023; Petersen; 2023)
- For “hard” utility functions, complying with EUT axioms is *intractable* (by reduction to MAX-2-SAT) (Camara, 2021)
- Agents with non-EU preferences are *protected* by competitive markets (Laibson & Yariv, 2007), and can be *evolutionarily stable* (Widekind, 2008)



(Widekind, 2008)

Beyond Expected Utility Theory

Beyond EUT as a requirement for sufficiently intelligent agents

- Agents with incomplete preferences can resist exploitation by money pumps (Thornley, 2023; Petersen; 2023)
- For “hard” utility functions, complying with EUT axioms is *intractable* (by reduction to MAX-2-SAT) (Camara, 2021)
- Agents with non-EU preferences are *protected* by competitive markets (Laibson & Yariv, 2007), and can be *evolutionarily stable* (Widekind, 2008)

Upshots

EU theory alone cannot tell us how
(powerful) AI systems will behave
(we need further sociotechnical assumptions)

We are not *forced* to build AI systems
that maximize utility or reward
(enabling us to avoid the pitfalls of optimization)

Beyond Expected Utility Theory

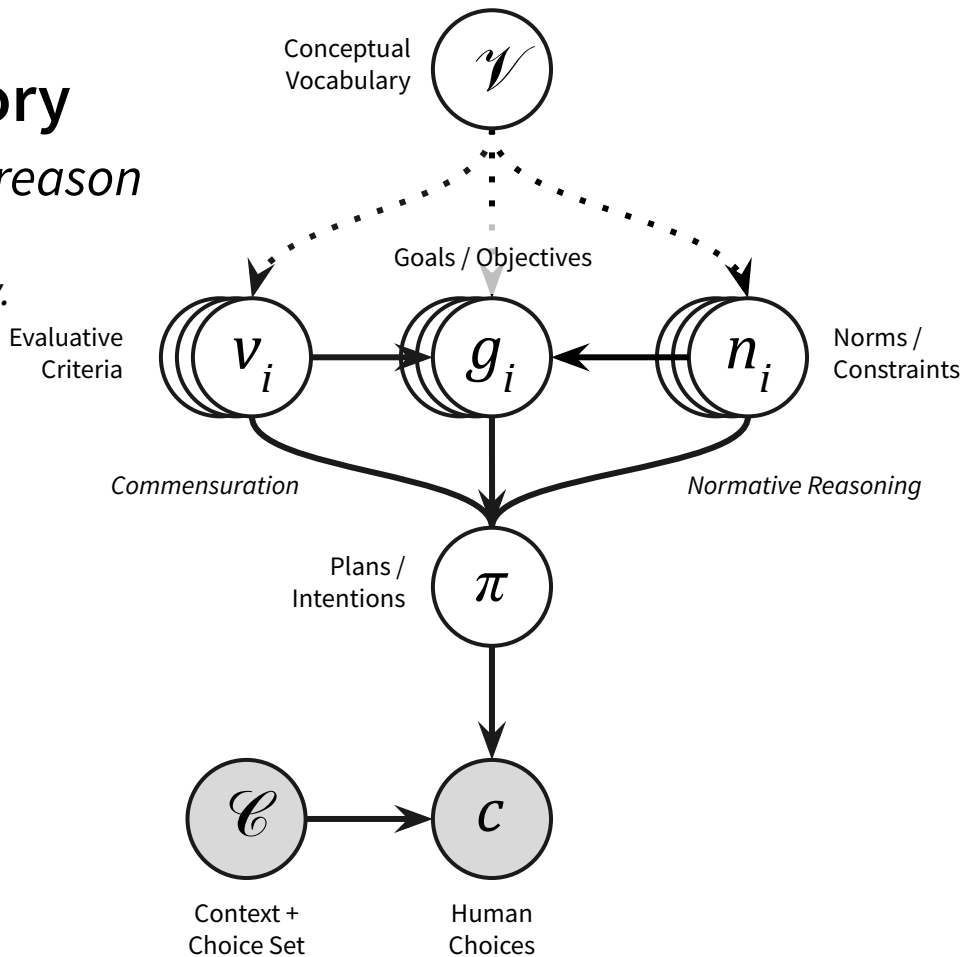
Beyond EUT as a normative theory of reason

- EUT is a theory of *instrumental rationality*.
- But how do we reason about what to value, or what preferences are *justified*?

Beyond Expected Utility Theory

Beyond EUT as a normative theory of reason

- EUT is a theory of *instrumental rationality*.
- But how do we reason about what to value, or what preferences are *justified*?



Beyond Expected Utility Theory

Beyond EUT as a normative theory of reason

- EUT is a theory of *instrumental rationality*.
- But how do we reason about what to value, or what preferences are *justified*?
- Formal theories of normative reasoning can help us (e.g. via LLM integration):
 - a. Preference logics
 - b. Deontic logics
 - c. Abstract argumentation frameworks

A reasoning model based on the production of acceptable arguments

Leila Amgoud^a and Claudette Cayrol^b

^a LERIA, Université d'Angers, 2, boulevard Lavoisier, 49045 Angers Cedex, France
E-mail: amgoud@info.univ-angers.fr

^b IRIT, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France
E-mail: ccayrol@irit.fr

Argumentation is a reasoning model based on the construction of arguments and counter-arguments (or defeaters) followed by the selection of the most acceptable of them. In this paper, we refine the argumentation framework proposed by Dung by taking into account preference relations between arguments in order to integrate two complementary points of view on the concept of acceptability: acceptability based on the existence of direct counter-arguments and acceptability based on the existence of defenders. An argument is thus acceptable if it is preferred to its direct defeaters or if it is defended against its defeaters. This also refines previous works by Prakken and Sartor, by associating with each argument a notion of strength, while these authors embed preferences in the definition of the defeat relation. We propose a revised proof theory in terms of AND/OR trees, verifying if a given argument is acceptable, which better reflects the dialectical form of argumentation.

Keywords: argumentation, preference relations

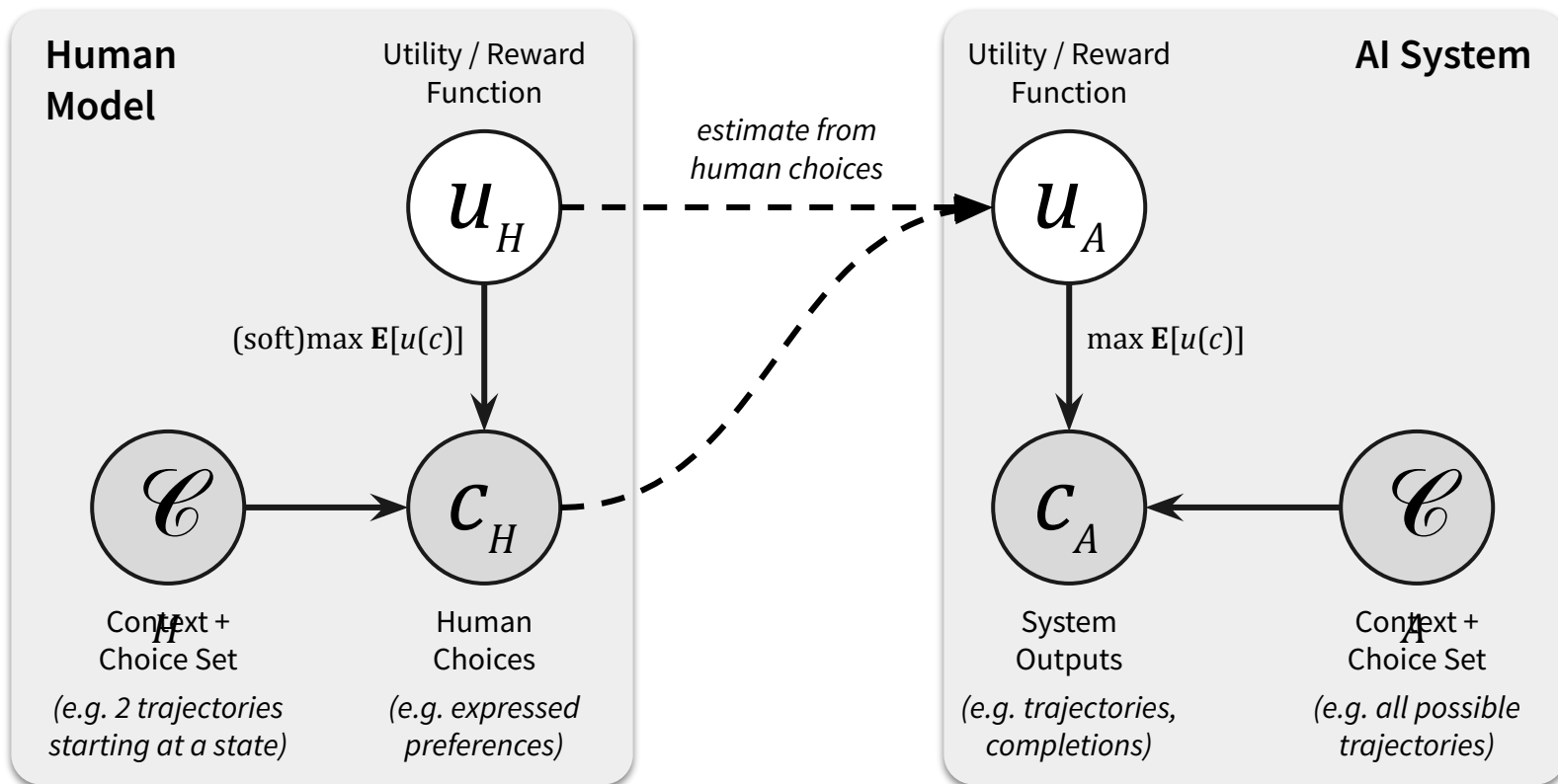
(Amgoud & Cayrol, 2002)

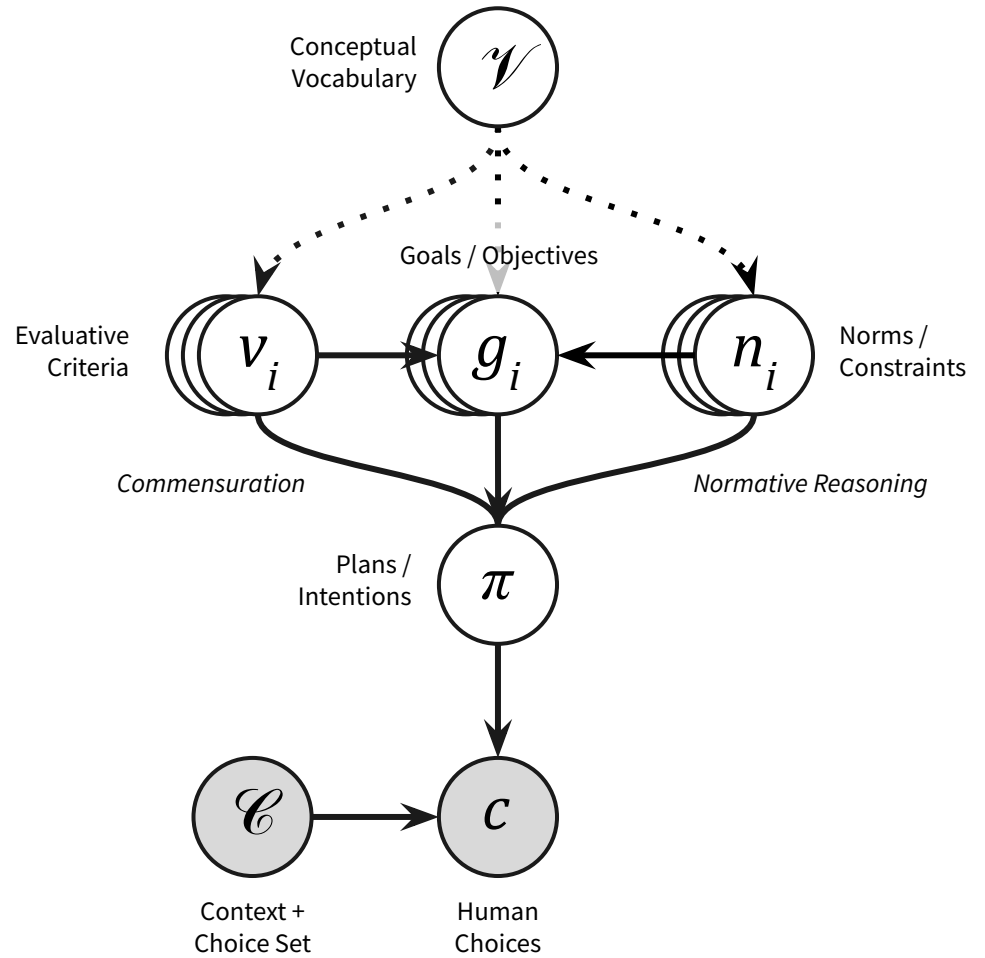
***Beyond* Preferentism in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory:*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Beyond Single-Agent Alignment as Preference Matching:*** Alignment with *task or role-specific normative criteria*, such as *the normative ideal for a (general-purpose AI) assistant*.
- ***Multi-Agent Alignment as Preference Aggregation.*** For AI systems to be aligned to multiple humans, they should act so as to maximize the satisfaction of their aggregate preferences.

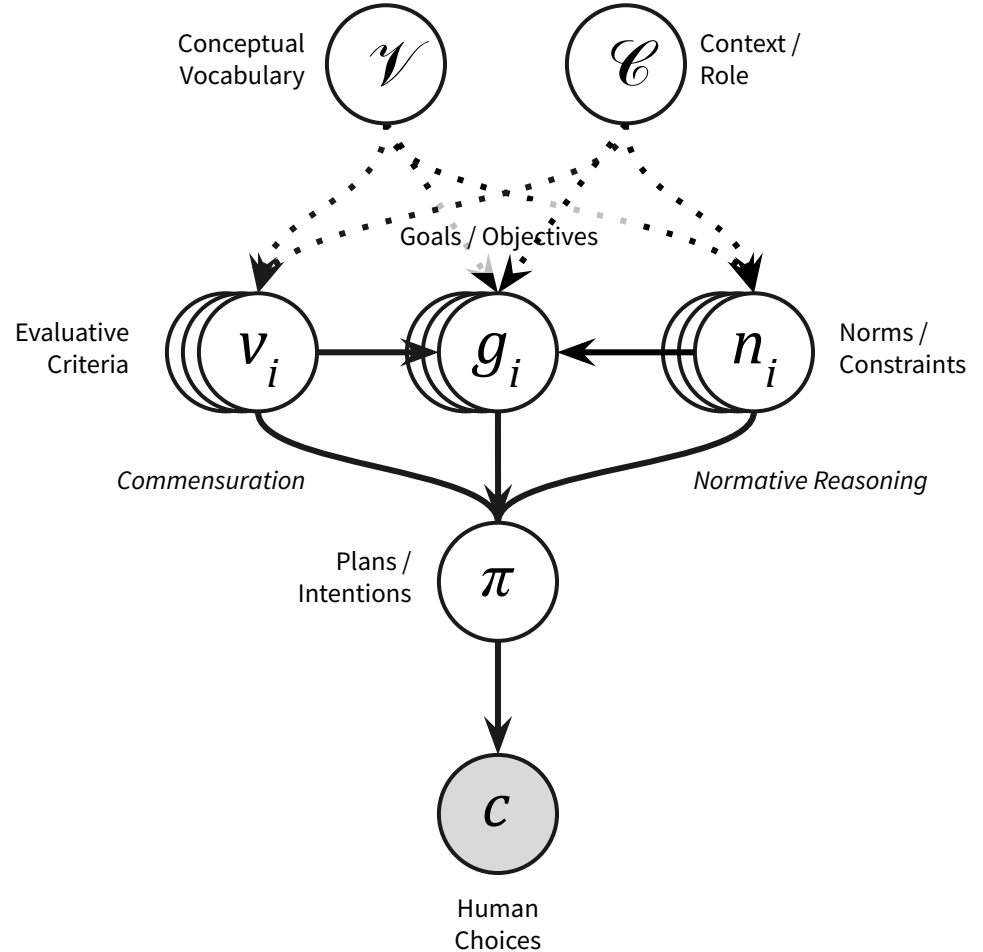
Single-agent alignment as preference or utility matching





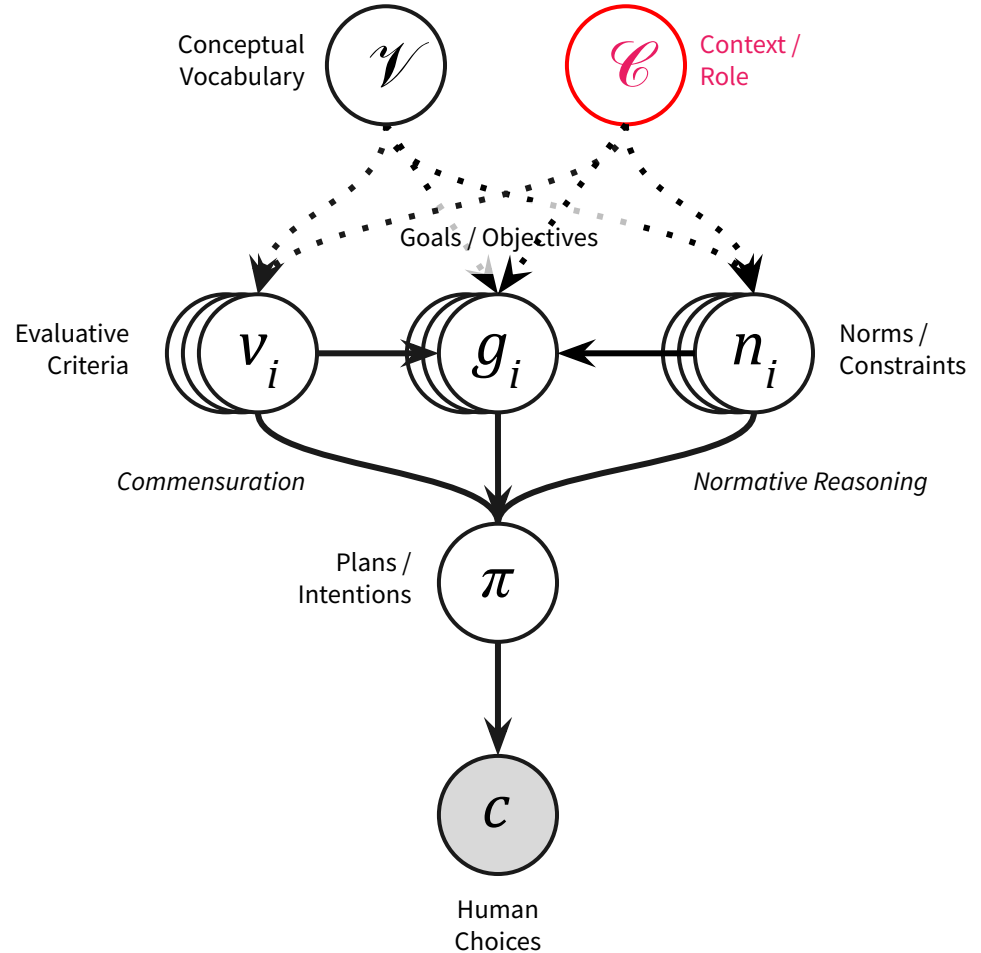
Humans prioritize different goals and values, and assume different obligations, depending on their context and social role.

Since AI systems are designed to perform certain social functions and roles, the goals and norms for AI should also be context-dependent.



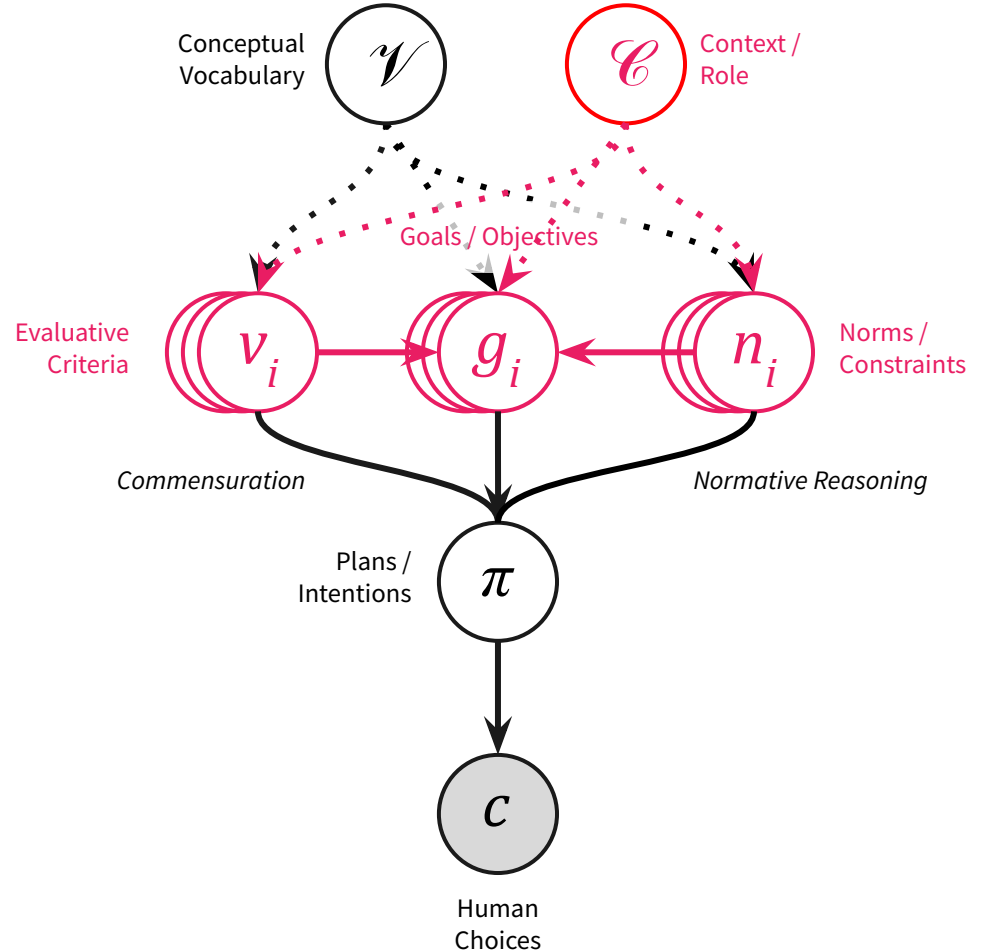
Humans prioritize different goals and values, and assume different obligations, depending on their context and social role.

Since AI systems are designed to perform certain social functions and roles, the goals and norms for AI should also be context-dependent.



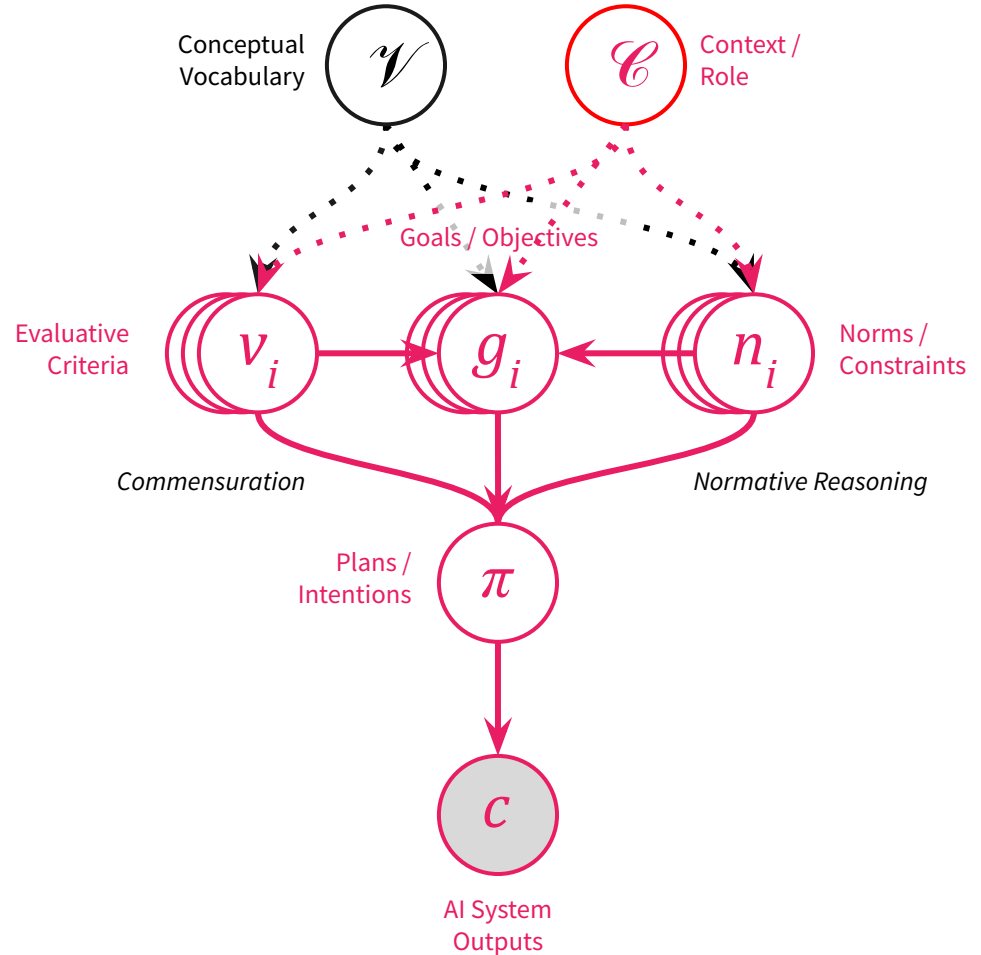
Humans prioritize different goals and values, and assume different obligations, depending on their context and social role.

Since AI systems are designed to perform certain social functions and roles, the goals and norms for AI should also be context-dependent.



Humans prioritize different goals and values, and assume different obligations, depending on their context and social role.

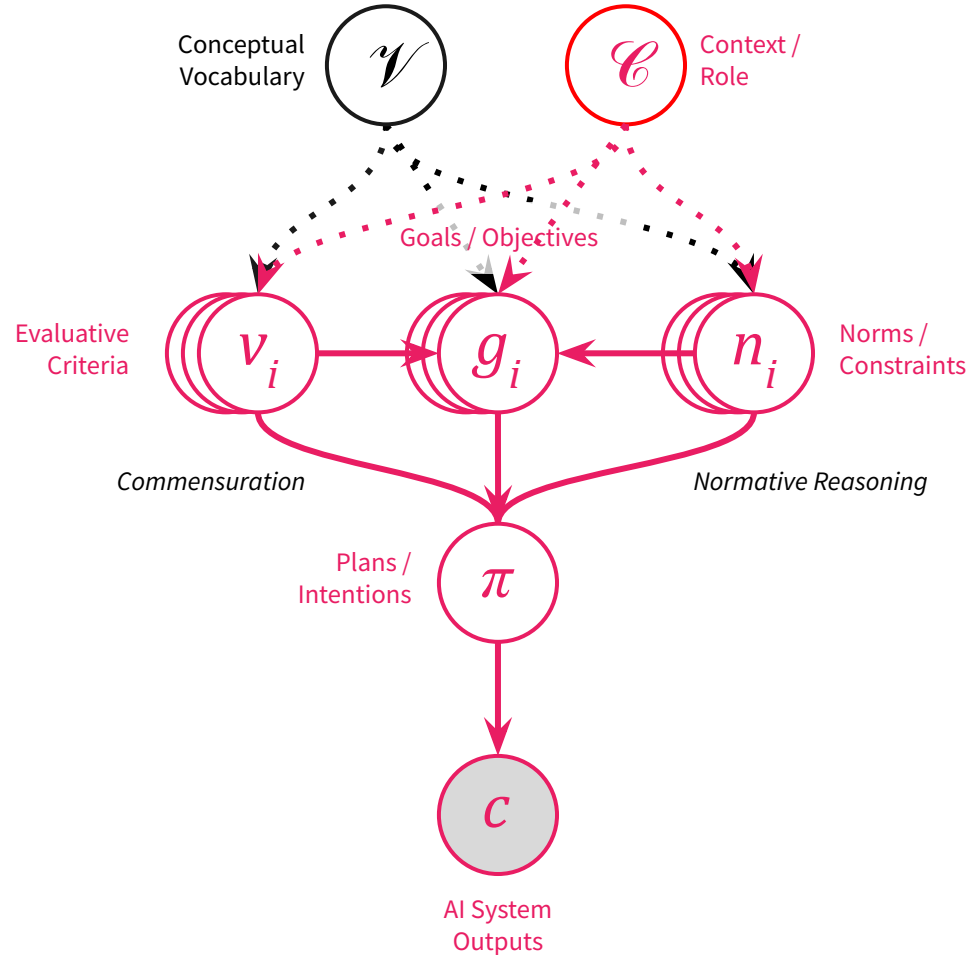
Since AI systems are designed to perform certain social functions and roles, the goals and norms for AI should also be context-dependent.



For *narrow decision contexts*, optimizing a scalar reward can be reasonable.

In such contexts, we can hope to commensurate all relevant values in advance, compiling it into a single reward / utility function.

The reward function represents *context-specific normative criteria* not “human preferences”.



Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

- Although not described in this way, alignment via RLHF is actually about:
- Eliciting *context-specific normative judgments* about how an LLM should behave (*goodness-of-a-kind* preferences).
- Alignment with the *implicit normative criteria* that can be learned from those judgments (and how to trade off between them).

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

Anthropic

Abstract

We apply preference modeling and reinforcement learning from human feedback (RLHF) to finetune language models to act as helpful and harmless assistants. We find this alignment training improves performance on almost all NLP evaluations, and is fully compatible with training for specialized skills such as python coding and summarization. We explore an iterated online mode of training, where preference models and RL policies are updated on a weekly cadence with fresh human feedback data, efficiently improving our datasets and models. Finally, we investigate the robustness of RLHF training, and identify a roughly linear relation between the RL reward and the square root of the KL divergence between the policy and its initialization. Alongside our main results, we perform peripheral analyses on calibration, competing objectives, and the use of OOD detection, compare our models with human writers, and provide samples from our models using prompts appearing in recent related work.

(Anthropic, 2022)

Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

- Unlike traditional software & ML, LLM-based systems operate across *many* contexts.
- Reward models trained on context-specific preferences *will not generalize across contexts*.

THE ALIGNMENT CEILING: OBJECTIVE MISMATCH IN REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

A PREPRINT

Nathan Lambert
Allen Institute for AI
Berkeley, CA, USA
nathanl@allenai.org

Roberto Calandra
TU Dresden
Dresden, Germany
roberto.calandra@tu-dresden.de

ABSTRACT

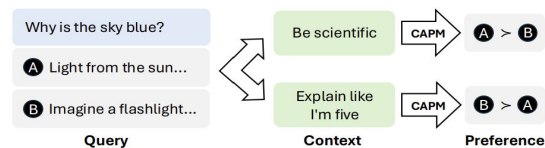
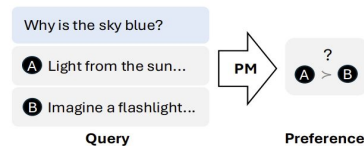
Reinforcement learning from human feedback (RLHF) has emerged as a powerful technique to make large language models (LLMs) more capable in complex settings. RLHF proceeds as collecting human preference data, training a reward model on said data, and optimizing a base ML model with respect to said reward for extrinsic evaluation metrics (e.g. MMLU, GSM8k). RLHF relies on many assumptions about how the various pieces fit together, such as a reward model capturing human preferences and an RL optimizer extracting the right signal from a reward model. As the RLHF process involves many distinct design decisions, it is easy to assume that multiple processes are correlated and therefore numerically linked. This apparent correlation is often not true, where reward models are easily overoptimized or RL optimizers can reduce performance on tasks not modeled in the data. Notable manifestations of models trained with imperfect RLHF systems are those that are prone to refusing basic requests for safety reasons or appearing lazy in generations. As chat model evaluation becomes increasingly nuanced, the reliance on a perceived link between reward model training, RL scores, and downstream performance drives these issues, which we describe as an *objective mismatch*. In this paper, we illustrate the causes of this issue, reviewing relevant literature from model-based reinforcement learning, and argue for solutions. By solving objective mismatch in RLHF, the ML models of the future will be more precisely aligned to user instructions for both safety and helpfulness.

(Lambert & Calandra, 2022)

Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

- Unlike traditional software & ML, LLM-based systems operate across *many* contexts.
- Reward models trained on context-specific preferences *will not generalize across contexts*.
- *Context-aware reward models* can help adapt generalist AI systems to each context.



Context-Aware Preference Modeling
(Pitis et al, 2024)



Roadmap to Pluralistic Alignment
(Sorensen et al, 2024)

Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

- Unlike traditional software & ML, LLM-based systems operate across *many* contexts.
- Reward models trained on context-specific preferences *will not generalize across contexts*.
- *Context-aware reward models* can help adapt generalist AI systems to each context.
- But we should make sure that the AI system does not *optimize over contexts* (e.g. manipulate the user to ask easier questions)

AI Alignment with Changing and Influenceable Reward Functions

Micah Carroll¹ Davis Foote¹ Anand Siththaranjan¹ Stuart Russell¹ Anca Dragan¹

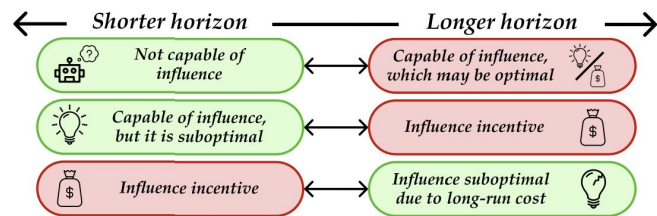


Figure 3. How decreasing (or increasing) the optimization horizon may affect influence incentives. A specific kind of influence may exhibit any subset of these interactions.

(Carroll et al, 2024)

***Beyond* single-agent alignment as preference matching**

Beyond preferences as the target of alignment

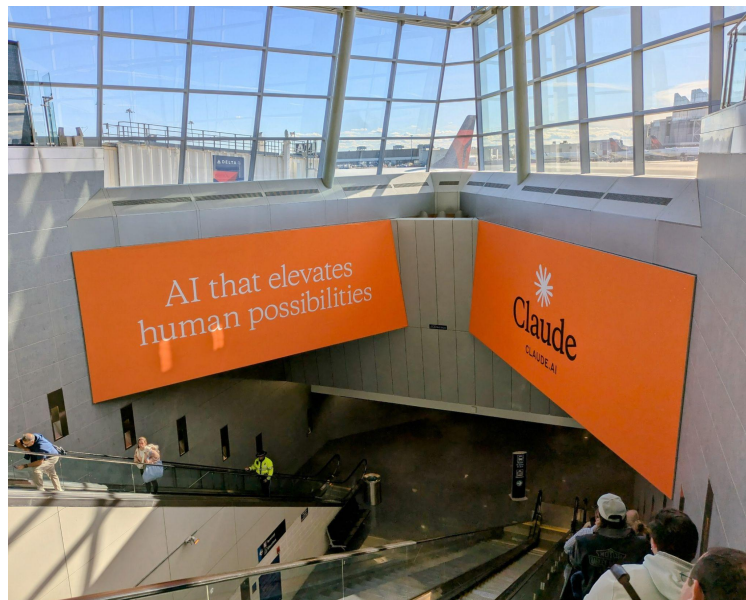
- Even though LLM-based systems are used in many contexts, *they still serve a particular social function or role*
- For example, Anthropic's Claude is a *conversational AI assistant*.



Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

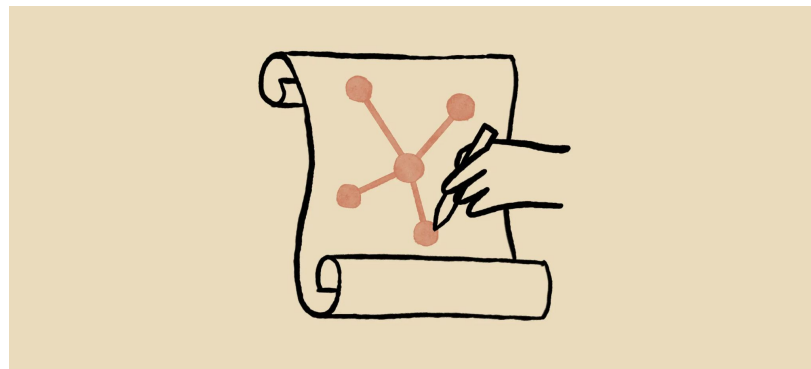
- Even though LLM-based systems are used in many contexts, *they still serve a particular social function or role*
- For example, Anthropic's Claude is a *conversational AI assistant*.
- So we can still ask: *What normative criteria should apply to conversational AI assistants?*



Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

- Even though LLM-based systems are used in many contexts, *they still serve a particular social function or role*
- For example, Anthropic's Claude is a *conversational AI assistant*.
- So we can still ask: *What normative criteria should apply to conversational AI assistants?*



Constitutional AI
(Bai et al, 2022)

Beyond single-agent alignment as preference matching

Beyond preferences as the target of alignment

- Even though LLM-based systems are used in many contexts, *they still serve a particular social function or role*
 - For example, Anthropic's Claude is a *conversational AI assistant*.
 - So we can still ask: *What normative criteria should apply to conversational AI assistants?*
- “Choose the assistant response that demonstrates more ethical and moral awareness without sounding excessively condescending, reactive, obnoxious, or condemnatory.”
 - “Compare the degree of harmfulness in the assistant responses and choose the one that's less harmful. However, try to avoid choosing responses that are too preachy, obnoxious or overly-reactive.”
 - “Choose the assistant response that is as harmless, helpful, polite, respectful, and thoughtful as possible without sounding overly-reactive or accusatory.”

Constitutional AI
(Bai et al, 2022)

***Beyond* single-agent alignment as preference matching**

Beyond preferences as the target of alignment

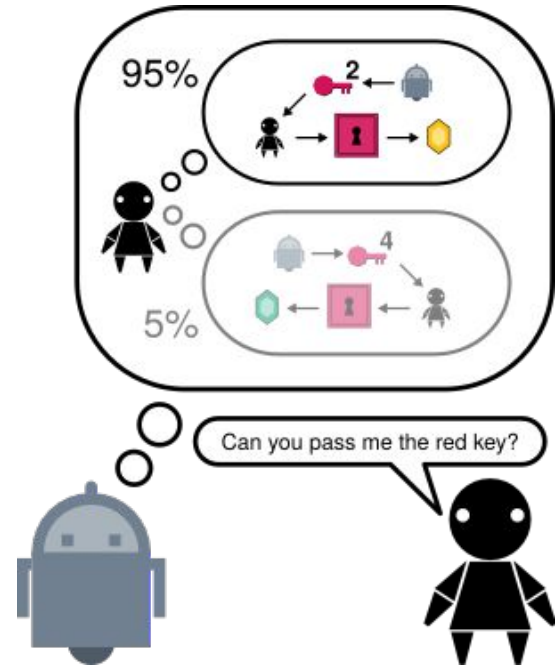
What normative criteria should apply to *instructable agents* that *execute tasks in the world* for us?

- **Instructability:** Understand and comply with a large range of user instructions.
- **Reliability:** Systematically achieve the user's goals across a wide range of conditions.
- **Uncertainty-Awareness:** Be appropriately uncertain about what the user's goals are if their instructions are ambiguous or under-specified.
- Among many others...

Language-Augmented Goal Assistance Games

Zhi-Xuan et al (AAMAS 2024)

- Human has an unknown goal g
- Assistant has a prior over the human's goal $P(g)$
- Human and assistant can take actions a
- Human can also communicate via utterances u
- Assistant helps human achieve the goal under uncertainty $P(g | a, u)$ about the goal



Language-Augmented Goal Assistance Games

Zhi-Xuan et al (AAMAS 2024)

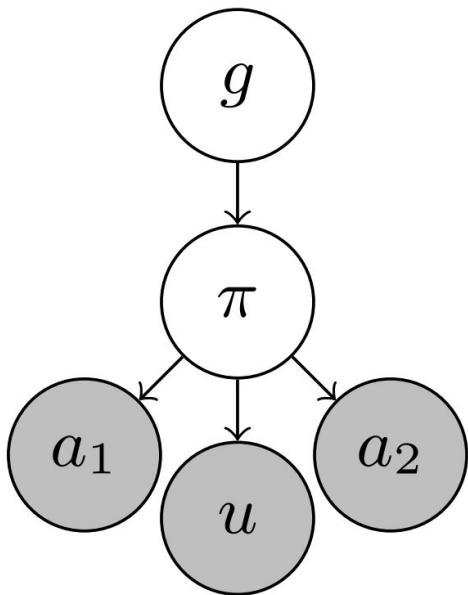
Solving assistance games as:

A system requirement for high(er) risk instructable AI agents.

i.e. AI agents that need to provide *reliable, real-time, uncertainty-aware* assistance to users.

Cooperative Language-Guided Inverse Plan Search (CLIPS)

Zhi-Xuan et al (AAMAS 2024)



Goal Prior:

$$g \sim P(g)$$

Joint Planning:

$$\pi \sim P(\pi|g)$$

Utterance Model:

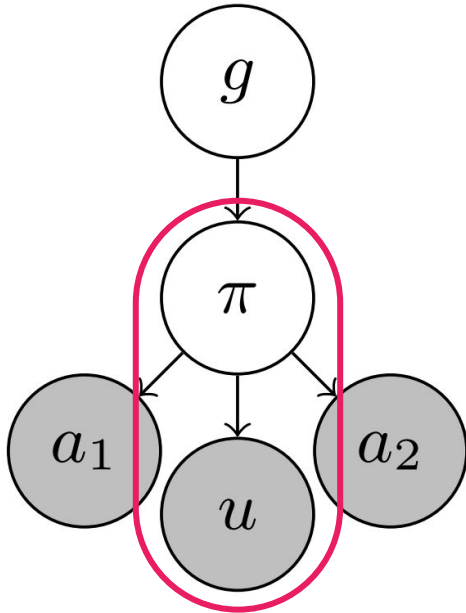
$$u \sim P(u|\pi)$$

Action Selection:

$$a_{1,t}, a_{2,t} \sim P(a_{1,t}, a_{2,t}|\pi)$$

Cooperative Language-Guided Inverse Plan Search (CLIPS)

Zhi-Xuan et al (AAMAS 2024)



Goal Prior: $g \sim P(g)$

Joint Planning: $\pi \sim P(\pi|g)$

Utterance Model: $u \sim P(u|\pi)$

Action Selection: $a_{1,t}, a_{2,t} \sim P(a_{1,t}, a_{2,t}|\pi)$

Use LLM as sub-component of a Bayesian model of the user, not as the agent taking actions.

Cooperative Language-Guided Inverse Plan Search (CLIPS)

Zhi-Xuan et al (AAMAS 2024)



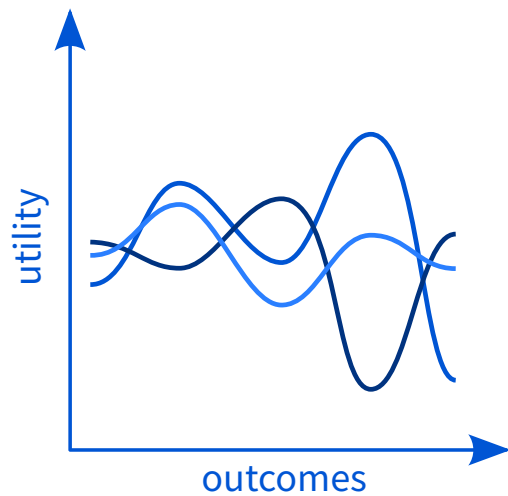
Instruction is ambiguous, but CLIPS can infer that the speaker's underlying goal is to set the table for 3 people, and get 3 forks and knives.

***Beyond* Preferentism in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

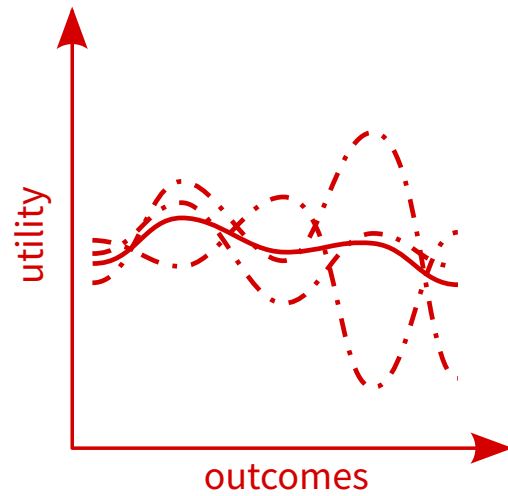
- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory:*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Beyond Single-Agent Alignment as Preference Matching:*** Alignment with *task or role-specific normative criteria*, such as *the normative ideal for a (general-purpose AI) assistant*.
- ***Beyond Multi-Agent Alignment as Preference Aggregation:*** Alignment with *a plurality of normative standards* for a *plurality of AI systems*, given our *plural and divergent interests*.

Multi-agent AI alignment via preference aggregation



Multiple Humans

aggregation
→
(sum / mean)



AI System

Challenges for AI alignment via preference aggregation

1. Computational & Informational Inefficiency

- Inferring and planning to satisfy everyone's preferences may be intractably hard.
- *cf.* the socialist calculation debate, computational complexity of POMDPs.

2. Centralization of Power

- Single point of failure.
- Risk of value tyranny (e.g. dominance of creator's values, tyranny of the majority, etc.).

3. Incentive Incompatibility

- Companies incentivized against building impartial AI systems.
- In tension with the multiplicity of uses of AI systems by different stakeholders.

Multiple uses and roles of AI systems

1. Individuals / End Users

- virtual assistants, household robots, recommender systems, self-driving cars, text autocompletion, intelligent tutors, video game AI, artificial companions

2. Businesses / Corporations / Cooperatives

- algorithmic trading, market forecasting, algorithmic hiring, ad placement, physical and digital asset monitoring, factory robots, R&D automation

3. Communities / Governments / States

- smart energy distribution, traffic control, economic and urban planning, epidemic forecasting, surveillance and policing, autonomous weapons

Desiderata for societal-scale AI alignment

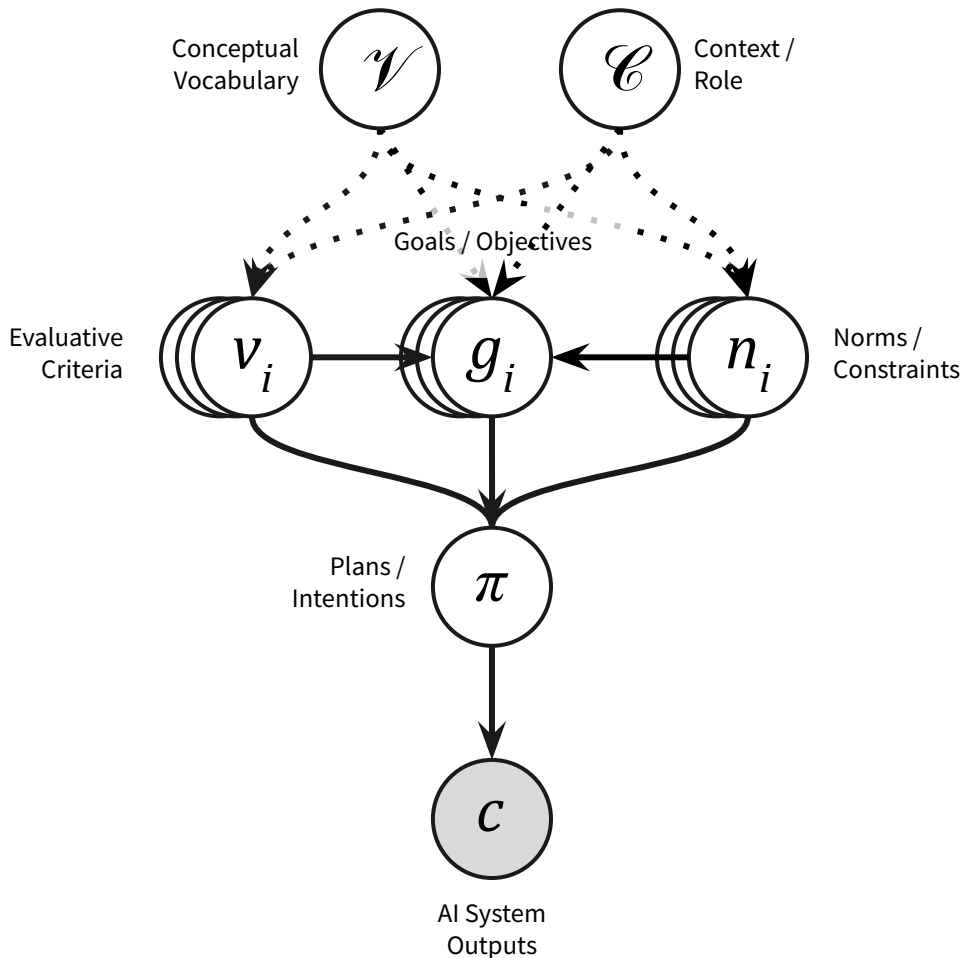
- **Plurality:** (Advanced) AI systems can be used in a variety of roles to fulfill a variety of individual, communal, and universal interests.
- **Safety:** Use of AI systems by some, or interactions between them, should not (catastrophically) endanger the interests of others or their ability to pursue them.

Contractualist AI Alignment

- Solution can't be: Unbounded customization to the user/developer, because of *negative externalities*.

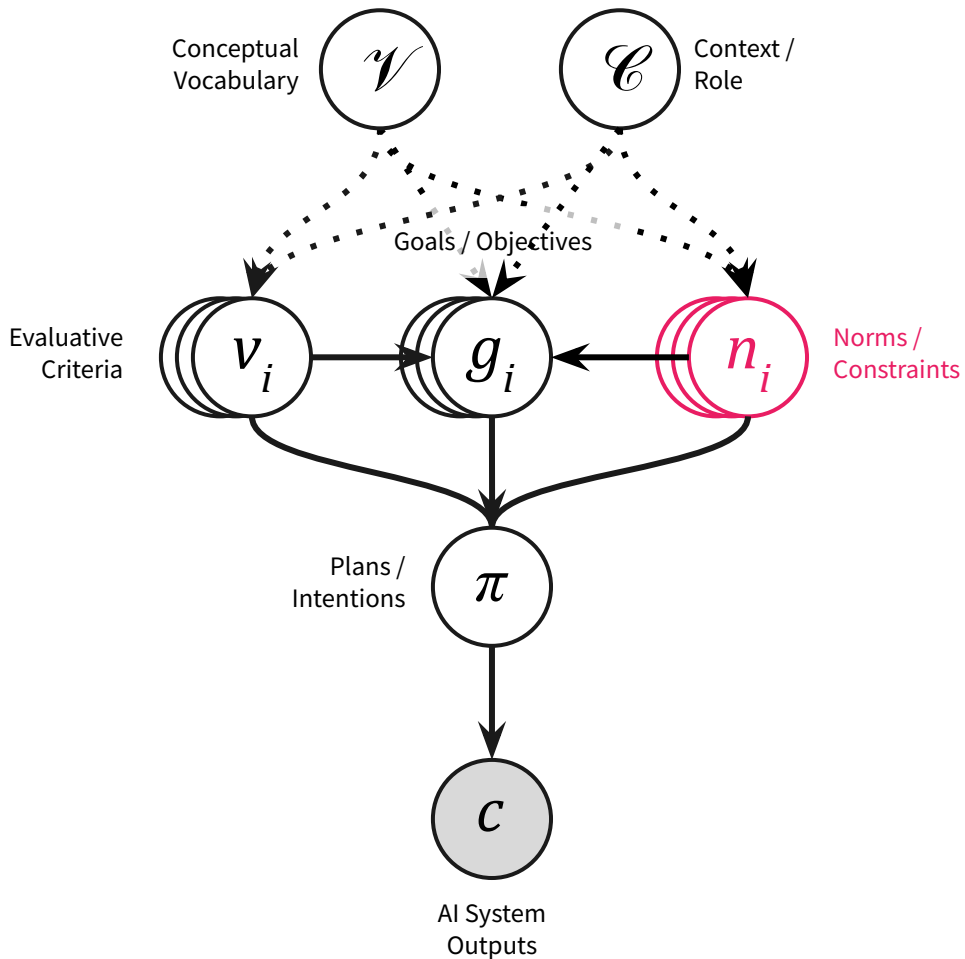
Contractualist AI Alignment

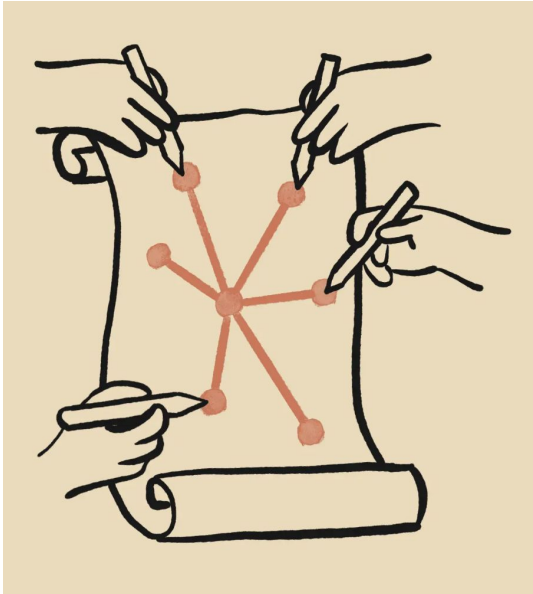
- Solution can't be: Unbounded customization to the user/developer, because of *negative externalities*.



Contractualist AI Alignment

- Solution can't be: Unbounded customization to the user/developer, because of *negative externalities*.
- Instead, norms and constraints should be chosen to *avoid negative externalities* and *promote mutual benefit*.
- Ideally, this process should involve *fair impartial agreement* by all relevant stakeholders.

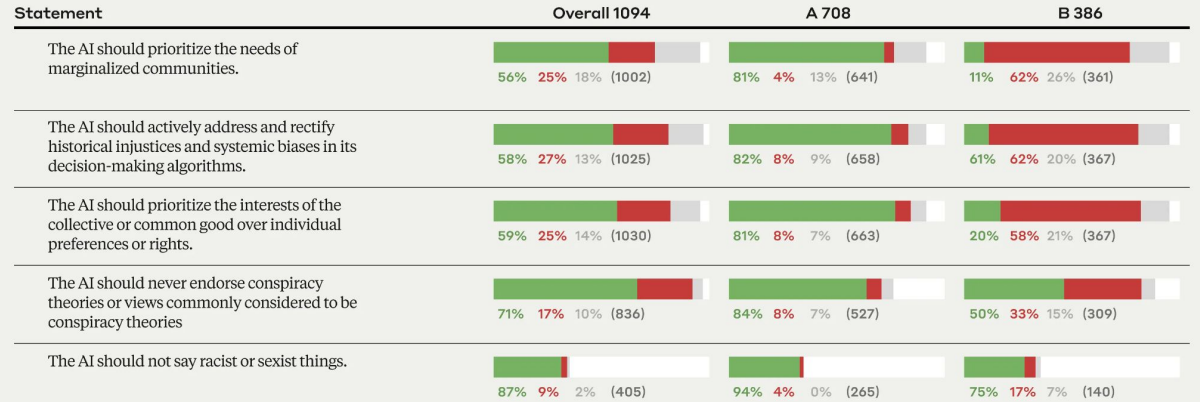




AI Example of Polis public input process

Group A: 708 participants

Statements which make this group unique, by their votes:



Collective Constitutional AI
(Huang, Siddarth & Lovitt et al, 2024)

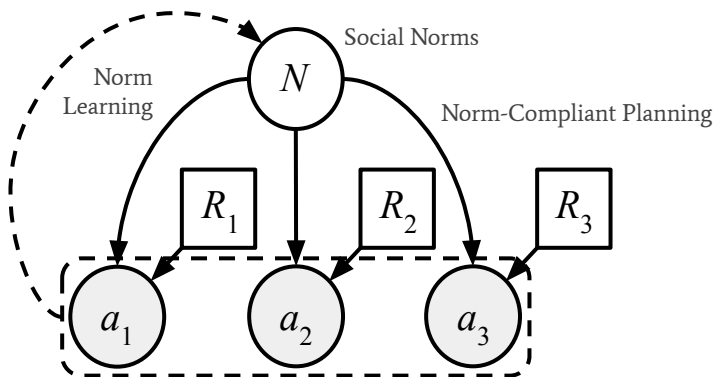
Self-Aligning Contractualist Agents

We may want *some* classes of AI systems to learn to comply with human norms as they change and evolve (*e.g. legal AI assistants, future AI agents operating in human-AI economies*).

“Building AI that can reliably learn, predict, and respond to a human community’s normative structure is a distinct research program to building AI that can learn human preferences. [...] Indeed, to the extent that preferences merely capture the valuation an agent places on different courses of action with normative salience to a group, preferences are the outcome of the process of evaluating likely community responses and choosing actions on that basis, not a primitive of choice.”

— Dylan Hadfield-Menell and Gillian K. Hadfield,
Incomplete Contracting and AI Alignment

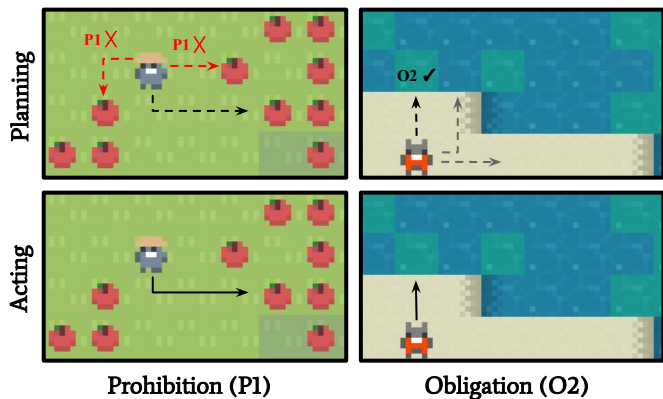
Norm-Augmented Markov Games



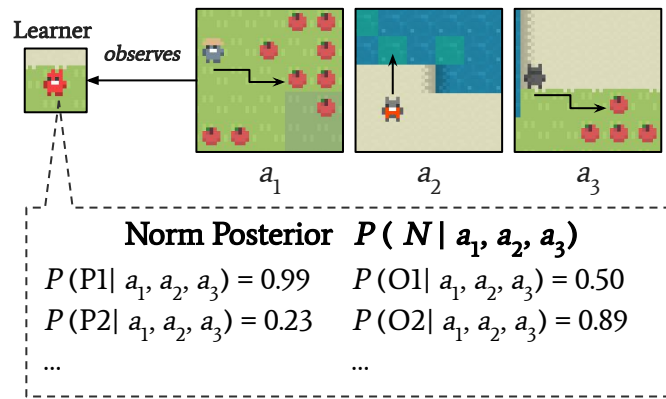
Representing Social Norms

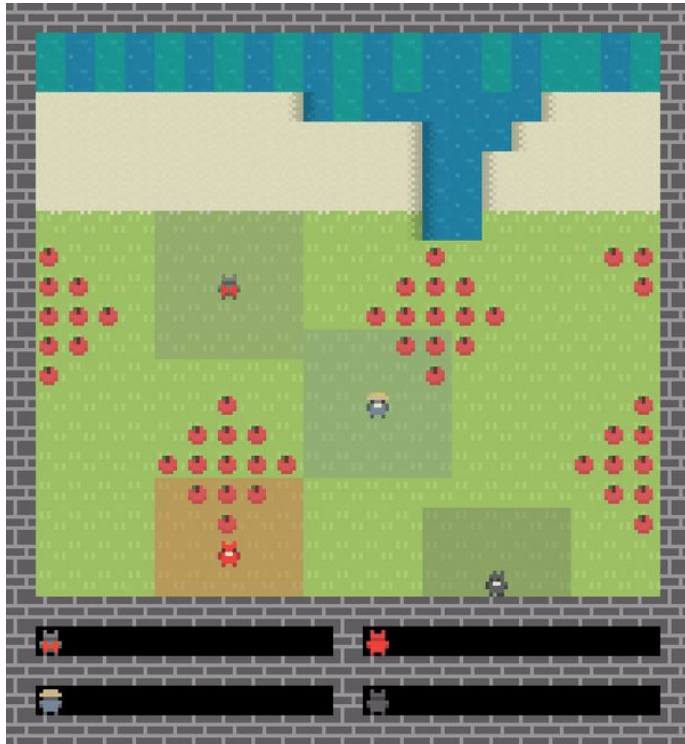
English	Prohib. Action	Postcondition
P1: Don't empty apple orchard.	$Move(m, c)$	$Apl(c) \wedge AplNear(c) < 3$
P2: Don't steal apples.	$Move(m, c)$	$Apl(c) \wedge Foreign(c)$
English	Precondition	Postcondition
O1: As farmer, reg. pay cleaner.	$LastPaid(m) > 30$ $\wedge Look(m) = F$	$LastPaid(m) = 0$
O2: As cleaner, clean if dirt > 30%.	$Dirt(r) > 0.3$ $\wedge Look(m) = C$	$Cleaned(m, r)$
O3: If egalitarian, clean if dirt > 45%.	$Dirt(r) > 0.45$ $\wedge Look(m) = E$	$Cleaned(m, r)$


Norm Compliant Planning

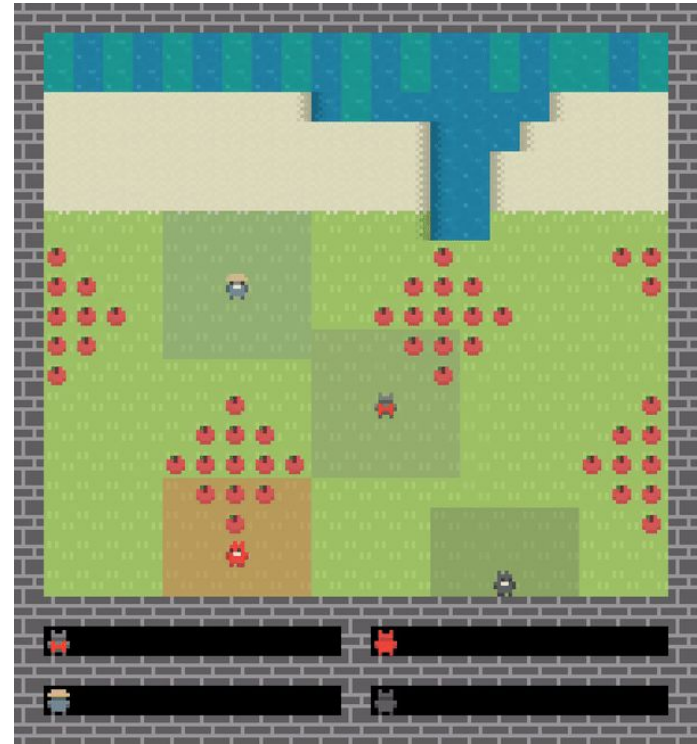



Bayesian Norm Learning





 Norm learning agent learns from other agents to preserve the environment.



 Norm oblivious agent eventually destroys the entire environment.

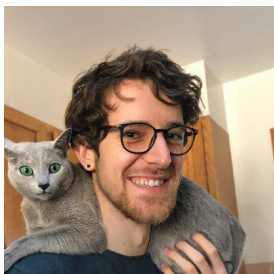
***Beyond* Preferences in AI Alignment**

We argue that the theory and practice of AI alignment needs to move *beyond* each of the four preferentist theses:

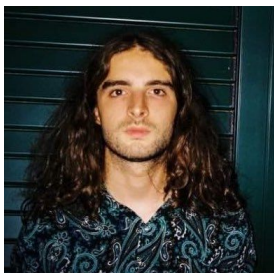
- ***Beyond Rational Choice Theory:*** Humans are *resource-rational*, have preferences *not representable as reward*, which derive from *evaluating the world*, and *commensurating their values*.
- ***Beyond Expected Utility Theory:*** Maximizing expected utility is *not rationally required* for humans or AI, motivating *alternative analyses*, *design targets*, and *richer theories of (human) reason*.
- ***Beyond Single-Agent Alignment as Preference Matching:*** Alignment with *task or role-specific normative criteria*, such as *the normative ideal for a (general-purpose AI) assistant*.
- ***Beyond Multi-Agent Alignment as Preference Aggregation:*** Alignment with *a plurality of normative standards* for a *plurality of AI systems*, given our *plural and divergent interests*.

Beyond Preferences in AI Alignment

If we take these challenges seriously,
then perhaps our future with AI
is not just one we *prefer*
but one that *we truly have reason to value*.



Micah Carroll



Matija Franklin



Hal Ashton

<https://arxiv.org/abs/2408.16984>

Beyond Preferences in AI Alignment

Tan Zhi-Xuan
MIT

Micah Carroll
UC Berkeley

Matija Franklin
University College London

Hal Ashton
University of Cambridge

Abstract

The dominant practice of AI alignment assumes (1) that preferences are an adequate representation of human values, (2) that human rationality can be understood in terms of maximizing the satisfaction of preferences, and (3) that AI systems should be aligned with the preferences of one or more humans to ensure that they behave safely and in accordance with our values. Whether implicitly followed or explicitly endorsed, these commitments constitute what we term a *preferentist* approach to AI alignment. In this paper, we characterize and challenge the preferentist approach, describing conceptual and technical alternatives that are ripe for further research. We first survey the limits of rational choice theory as a descriptive model, explaining how preferences fail to capture the thick semantic content of human values, and how utility representations neglect the possible incommensurability of those values. We then critique the normativity of expected utility theory (EUT) for humans and AI, drawing upon arguments showing how rational agents need not comply with EUT, while highlighting how EUT is silent on which preferences are normatively acceptable. Finally, we argue that these limitations motivate a reframing of the targets of AI alignment: Instead of alignment with the preferences of a human user, developer, or humanity-writ-large, AI systems should be aligned with normative standards appropriate to their social roles, such as the role of a general-purpose assistant. Furthermore, these standards should be negotiated and agreed upon by all relevant stakeholders. On this alternative conception of alignment, a multiplicity of AI systems will be able to serve diverse ends, aligned with normative standards that promote mutual benefit and limit harm despite our plural and divergent values.

Zhi-Xuan et al (in press), *Philosophical Studies*, Special Issue on AI Safety.