Building Human-AI Alignment
Specifying, Inspecting, Modeling, and Revising AI Behaviors

serena.l.booth@gmail.com
slbooth.com

I investigate critical and timeless questions in artificial intelligence (AI): how can we as humans write specifications that meet our objectives, how can we build systems that learn from those specifications, and how can we know that the behaviors of those systems are aligned to our specifications? The field of AI suffers from rampant underspecification, overspecification, and misspecification, creating a grand challenge to safely align AI with human goals. This challenge spans the remit from niche AI tasks to large foundation AI models that will affect billions of people. It will continue to exist as more AI models are deployed throughout society: alignment problems will reach every sector of the economy.

Writing specifications for AI systems is critical yet notoriously hard because these systems lack common sense reasoning, making it easy to write specifications that result in unintended and potentially dangerous side effects. I have studied how AI experts often write erroneous specifications; for example, in one of my studies, over half of *experts* wrote erroneous specifications for a trivial setting. I am currently designing mechanisms to prevent experts from making these errors. In systems like ChatGPT, the AI learns a specification from non-experts' preferences over AI outputs. Here, too, my studies have shown that non-experts' preferences are often misinterpreted. I have identified how assumptions about human preferences can be flawed, and how this process should be revised to prevent misinterpretation. Humans must also learn about the capabilities and limitations of AI decision-making—these must be transparent so that humans can assess when and when not to rely on an AI system. In service of this requirement, I have designed methods that expose information-rich examples of AI system behaviors to humans, and I have studied cognitive science theories on how humans learn new concepts to let them better comprehend AI systems. I have shown how my methods can be used to revise specifications.

In service of human-AI alignment, I develop new mathematical models and algorithms, and conduct empirical analyses. I am a reinforcement learning researcher with expertise in learning from human feedback. I use Bayesian inference methods [6, 24] and learn models from human data [16, 15], as both approaches can help model human decision-making in the face of uncertainty. I run large-scale computational experiments to inspect and assess the correctness of hypotheses and claims about learning systems [2, 25, 23]. I regularly design human studies [2, 3, 5, 12] as simulated models of users are insufficient to assess human-AI interactions, and I use qualitative research methods like thematic analysis to understand both the outcomes of human decision-making and the reasoning process [2, 11]. I often consider robotics as an application.

Given the impact of new AI tools and their errors, governments around the world are scrambling to decide whether and how to establish guardrails in their development and use. In the U.S., the Biden Administration issued an executive order to this effect. I am currently working as a AAAS AI Policy Advisor in the U.S. Senate to implement this order for the Banking, Housing, and Urban Affairs committee. I have also co-authored an IEEE standard for autonomous systems that defines transparency as a measurable and testable property for safety certification agencies to assess AI systems [20, 21].

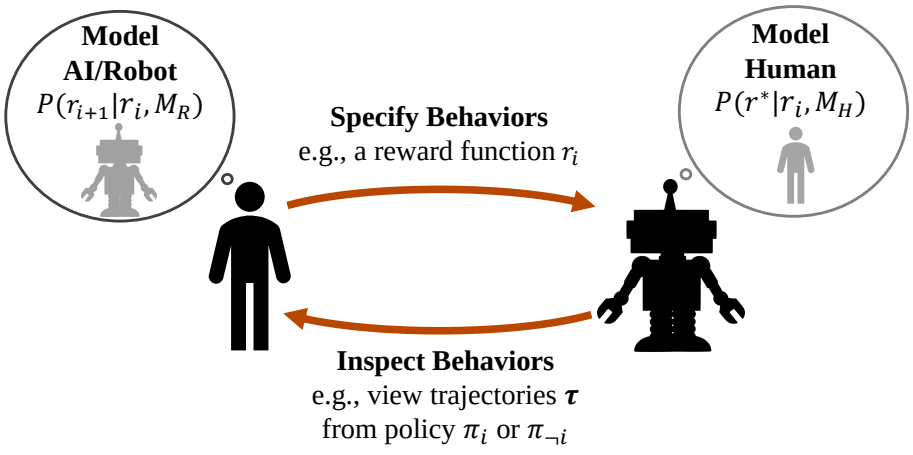


Figure 1: A sketch of my research. A human specifies an objective with a reward function r_i [2] or other signal like a dataset D_i of preferences [16, 15]. The robot models the human M_H to interpret the reward function, e.g., inferring a new reward function r^* given r_i and M_H. The robot learns a behavioral policy π_i, then presents expository examples of its behavior to help the human inspect it, for example as trajectories [6, 24, 3]. The human updates their conceptual model of the robot (M_R) [4] and the reward function r_i iteratively.

Specifying Behaviors through Reward Functions

Reinforcement learning (RL) is a promising approach for building robot and AI systems. Reward functions are an exceptionally flexible framework for specifying behaviors, and, as such, there is tremendous optimism about RL's potential, with some researchers even arguing that reward can specify the nature of intelligence [19]. RL's usefulness is nonetheless limited by the difficulty of specifying the reward function, which can be misspecified or underspecified [1]. While there is a push toward learning reward functions, manually-specified reward functions remain commonplace and have recently been used to achieve tremendous successes—for example, to outrace human champions in Gran Turismo [22] and drone competitions [13]. I have studied human reward design processes [2]. I first showed that reward functions can be easily *overfit* to learning algorithms, wherein the reward function is overloaded to both encode the desired behavior and also facilitate fast and successful learning for a specific algorithm or hyperparameter choice, at the expense of interoperability and generality. Through a user study, I confirmed that this problem of overfitting equally manifests with human experts designing the reward functions. While I confirmed this problem of overfitting is indeed persistent, I was also surprised to discover that—in a trivial gridworld environment, the type of environment you might encounter in an RL101 class—the *majority of expert humans* wrote reward functions that failed to encode the task. I attribute these failures to the mismatched interpretations of the reward function between the human designers and the goals of RL algorithms writ large. Humans view reward functions with a myopic lens, as a mechanism for encoding the relative goodness of each possible state, but this differs from the RL objective of maximizing summed and discounted reward. This study raises the questions: how can we enable humans to write better reward functions, and how can we enable AIs to better interpret flawed reward functions?

Specifying Behaviors through Preferences

A complementary approach for crafting specifications for RL is to learn a reward function from non-expert human feedback like preferences. This approach has recently seen wide success as the key innovation in ChatGPT [18, 8]. My research shows that the inductive biases we place on learning reward functions from such data must be reconsidered if we are to learn correct reward functions. In the common implementations of RL from human feedback (RLHF), there is an underlying assumption that preferences are determined by a segment's partial return: the summed discounted reward over a segment. As Figure 2 shows, this assumption is patently false. My collaborators and I propose an alternative inductive bias: that human preferences are determined by a segment's regret, a measure of deviation from optimality [16]. We validate this new framing through a user study, computational experiments, and theoretical proof. We show that underlying reward functions cannot be generally recovered using the partial return model, while they can with our proposed regret model. We also collect a dataset of human preferences and find that these preferences are better explained by our proposed regret model. We subsequently study the consequences of embedding this incorrect assumption in learning systems [15]. This shows the importance of accurately modeling human decision-making and of designing correct inductive biases when interpreting noisy data. There are implications for manual reward function design, too: I plan to treat the manually-designed reward function as an observation about the intended reward function, similar to Hadfield-Menell et al.'s *inverse reward design* [10]. In such a system, we must proceed with caution in the design of the prior encoding common specification errors.

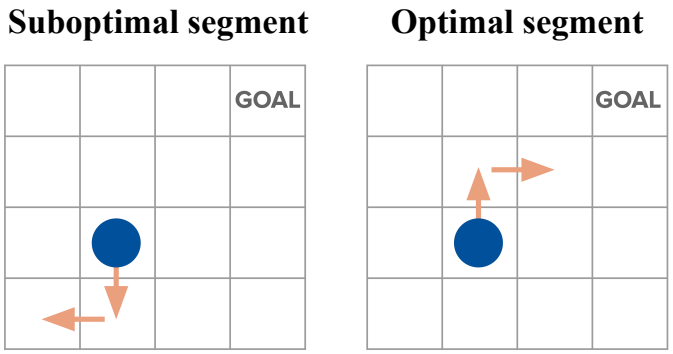
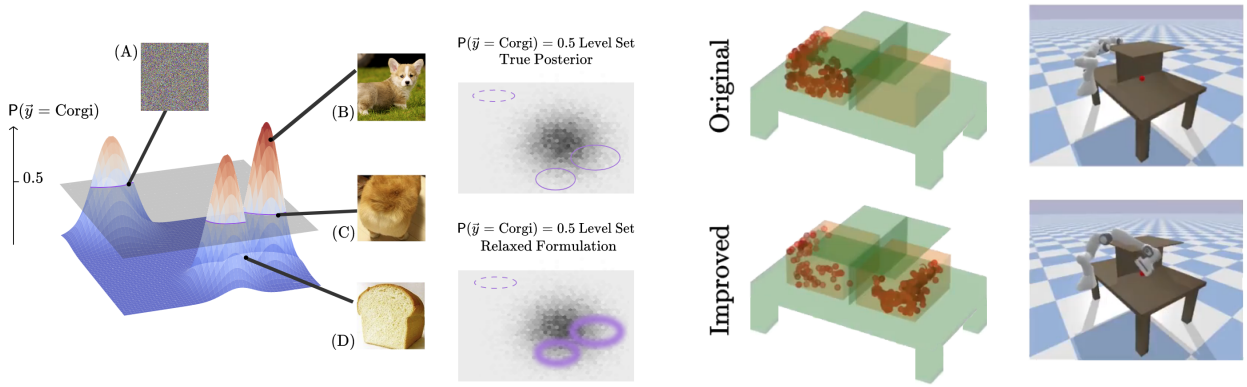


Figure 2: Modeling human decision-making priors is critical for alignment, yet popular methods make significant mistakes. Current RL from human feedback approaches consider both of these trajectories to be equally good, though intuitively the right is better. These approaches assume that human preferences are determined by *partial return*, the summed discounted reward [8, 18]. This assumption is wrong: a segment that navigates away from a goal has equal partial return to one that navigates toward the goal. We show that *regret*—a measure of suboptimality—is a better model of human preference that permits better inference of the intended reward function [16].



(a) Inspect Classifier Behaviors [6]

(b) Inspect Robot Behaviors [24]

Figure 3: AI behaviors must be inspected, e.g., by viewing examples, but finding informative examples is difficult. I introduce a method to do this by importance sampling the posterior of generative models [6, 24]. In Figure 3a, we find a 50% Corgi and 50% Bread example (C) to understand a decision boundary. In Figure 3b, we inspect a robot controller. We find the robot cannot reach the target red ball on the right side of the table (top). Once inspected, we find that increasing the collision clearance of the table's divider leads to success in reaching targets on either side of the divider (bottom).

Inspecting Behaviors

After writing a specification and using some algorithm to optimize it, how can a person assess whether a robot or an AI has learned the behavior that meets their needs and expectations? Is it aligned to their intent? The most common practice is to observe examples of the robot acting in the world in random or a fixed set of environments. However, without adding structure and discipline to this practice, the observation process is limited in usefulness. I propose that we instead support humans in searching for examples that communicate specific targeted behaviors. I first introduced a method for inspecting the behaviors of neural network or other classifiers. In Bayes-TrEx [6], a user specifies a prediction target (e.g., ambiguous across two classes) and a generative model, and then Bayesian inference finds examples that meet the prediction target. Bayes-TrEx helps with debugging and understanding neural networks, as it can be used to find *ambiguous* examples to communicate class boundaries or highly confident incorrect classifications to communicate systematic failures. We subsequently adapted this approach to create RoCUS, a method for debugging and improving robot controller behaviors by finding environments in which interesting behavior occurs [24]. We demonstrated that RoCUS can find and revise bugs in specifications in a dynamical system setting. RoCUS can also be used to assess the behaviors learned through RL with a reward function specification; this method can be applied to help the human iterate on their reward function design.

Building Conceptual Models

How do humans come to understand the behavioral patterns encoded in a reward function, or learned through a reward function? More generally, how do humans maintain and mitigate uncertainty about their beliefs about AI systems? This uncertainty relates to the human ability to form conceptual models, which are abstract models used for reasoning. Classroom-tested theories of human concept learning from the learning sciences community provide a rough blueprint for how to help people build and update accurate and flexible conceptual models [17, 9], and I show how these theories can be leveraged for human-robot interaction [4, 12]. These theories assert that conceptual models are best formed by experiencing examples that follow highly structured patterns of variance and invariance [17], and by experiencing structurally-aligned analogous examples that support rapid knowledge transfer [9]. When interacting with a robot or an AI system, a person will inevitably develop a conceptual model of the system's behaviors. But, without structure to their learning, the resulting conceptual model may be incorrect or inflexible. I have studied how these theories of human concept learning should be adapted for human-robot interaction: my analysis of 35 works showed ad-hoc incorporation of *some* of these patterns [4], but that the community still has many blind spots. For example, it is exceedingly rare to show counterexamples of robot capabilities, but counterexamples are essential for establishing the bounds of capabilities. My preliminary follow-up work confirms that viewing counterexamples of robot behaviors improves conceptual model formation [12]. My work provides design guidance for better structuring human observations of both AI system and robot learned behaviors.

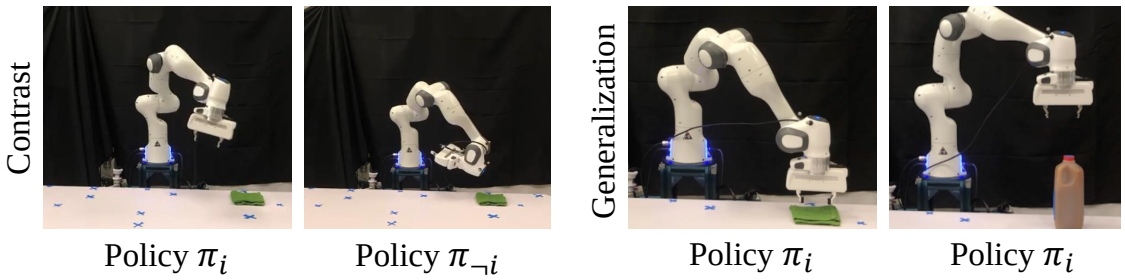


Figure 4: Humans must correctly model behaviors of AI agents like robots such that they can effectively collaborate with the agent and potentially revise an incorrect specification [4, 12]. My research shows that this modeling can be achieved efficiently by experiencing *contrast*: an alternative robot policy that delineates currently specified behaviors from alternatives (left). Experiencing contrast first helps the human to generalize behavior predictions to new settings (right).

Future Directions

How can we better interpret erroneous specifications and prevent humans from making errors? Specifications for AI systems will inevitably suffer from errors. By analyzing common pitfalls and errors made by both experts and non-experts, I aim to develop mechanisms that compensate for these mistakes by informing the inference of the intended specification. This line of proposed work would build upon Hadfield-Menell et al.'s inverse reward design, wherein each given reward function is viewed as an observation about the intended reward function [10]. Alongside compensating for errors in specifications, my research aims to prevent humans from making errors in the first place. This requires thinking about AI specification best practices—for example, researchers have shown benefits from considering the expected return of optimal and suboptimal trajectories and ensuring the ordering of preferences of these trajectories match the human's intentions [14].

How can we assess the alignment of AI systems? Even if a specification is aligned to human intent, at least with current reinforcement learning algorithms, there is no guarantee that the learned behavioral policy is perfectly optimized. We need to design methods to assess whether the learned policy expresses the same preferences as are encoded in the specification—to test the alignment of the system (e.g., [7]). For instance, the research community (me included) has built a set of tools to analyze AI behaviors (exemplars, critical states, policy summaries, probes, etc.) but this limited toolset may not yet allow humans to assess alignment, and true progress must judge their usefulness against the higher bar of verification.

How can we design AI systems to be governable? Currently, a leading method for governing large AI systems is RLHF. This training step allows us to establish safeguards by preventing these systems from expressing undesirable behaviors, but it is weak and can be 'jailbroken'. In my role as an AI Policy Advisor for the Senate Banking, Housing, and Urban Affairs committee, I oversee financial services. Given their fiduciary responsibility, financial service providers would be well-advised not to deploy a ChatBot that offered financial advice. RLHF does not guarantee that constraints cannot be undermined and finetuning largely erases constraints. I wish to design mechanisms to govern the behaviors of AI systems.

References

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

[2] Serena Booth, W Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5, pages 5920–5929, 2023.

[3] Serena Booth, Christian Muise, and Julie Shah. Evaluating the interpretability of the knowledge compilation map: Communicating logical statements effectively. In *IJCAI*, pages 5801–5807, 2019.

[4] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L Glassman. Revisiting human-robot teaching and learning through the lens of human concept learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 147–156. IEEE, 2022.

[5] Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, and Radhika Nagpal. Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 426–434, 2017.

[6] Serena Booth, Yilun Zhou, Ankit Shah, and Julie Shah. Bayes-TrEx: a Bayesian sampling approach to model transparency by example. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11423–11432, 2021.

[7] Daniel S Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. Value alignment verification. In *International Conference on Machine Learning*, pages 1105–1115. PMLR, 2021.

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in neural information processing systems*, pages 4299–4307, 2017.

[9] Dedre Gentner and Linsey A Smith. Analogical learning and reasoning. *The Oxford Handbook of Cognitive Psychology*, 2013.

[10] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in neural information processing systems*, 30, 2017.

[11] Aspen Hopkins and Serena Booth. Machine learning practices outside Big Tech: How resource constraints challenge responsible development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 134–145, 2021.

[12] Tiffany Horter, Elena L Glassman, Julie Shah, and Serena Booth. Varying how we teach: Adding contrast helps humans learn about robot motions. *HRI Human-Interactive Robot Learning (HIRL) Workshop*, 2023.

[13] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

[14] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023.

[15] W Bradley Knox, Stephane Hatgis-Kessell, Sigurdur Orn Adalgeirsson, Serena Booth, Anca Dragan, Peter Stone, and Scott Niekum. Learning optimal advantage from preferences and mistaking it for reward. *AAAI Conference on Artificial Intelligence*, 2023.

[16] W Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro Allievi. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*, 2022.

[17] Ference Marton. *Necessary conditions of learning*. Routledge, 2014.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[19] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[20] Alan Winfield, Eleanor Watson, Takashi Egawa, Emily Barwell, Iain Barclay, Serena Booth, Louise A Dennis, Helen Hastie, Ali Hossaini, Naomi Jacobs, et al. IEEE standard for transparency of autonomous systems. *IEEE*, 2022.

[21] Alan FT Winfield, Serena Booth, Louise A Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I Muttram, Joanna I Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, et al. IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, 8:665729, 2021.

[22] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[23] Yiming Zheng, Serena Booth, Julie Shah, and Yilun Zhou. The irrationality of neural rationale models. *arXiv preprint arXiv:2110.07550*, 2021.

[24] Yilun Zhou, Serena Booth, Nadia Figueroa, and Julie Shah. RoCUS: Robot controller understanding via sampling. *Conference on Robot Learning*, 2021.

[25] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.