Normative Agency Design and Evaluation

Silviu Pitis, November 2024

My main research interest lies in understanding and designing the goals of general-purpose, artificially-intelligent agents that serve humanity. What objectives should generalist AI agents pursue, and how can we evaluate their success? What does it mean for an agent to be "aligned" with humanity?

More specifically, my work focuses on sequential decision-making agents that pursue multiple objectives, and applies tools from language modeling, reinforcement learning, decision theory, social choice, causal modeling, and deep learning. This research statement motivates a normative view of alignment, and describes the three approaches that characterize my work: axiomatic reasoning about goal representations, leveraging structure to reason about goals, and most recently, expressing goals and evaluating agents using natural language.

1 Normative Alignment

When describing what we want from an AI agent, we often appeal to the empirical human, either as a model of desired behaviors, or as a principal that expresses preferences over possible behaviors [4, 6, 9, 11]. An agent that does what we want is said to be "aligned". Of course, this presupposes that we, as humans, know what we want. While this may be the case in specific circumstances, such as games with well-defined winning conditions, individual humans are often ambivalent, uncertain, or wrong. And matters become even less clear as the generality of agents increases and their actions begin to impact many humans, all at once. Indeed, humans disagree on a striking range of issues, many of which all sides view as fundamentally important. Even on simple matters, like choosing which of two responses by an AI chatbot is preferred, empirical human agreement is as low as 65% [7,26].

To elevate our study of agency beyond empirical humans, I believe we should view humans not as the targets of alignment, but as proxies to an underlying "ideal". What purposes would an ideal agent pursue? Or alternatively, what objectives would an ideal human (or society) give to its agents? To reason about a complex ideal that is, in many ways, super-human, it seems necessary to decompose the problem into smaller, more human-digestible parts. The approaches described below do this in distinct but complementary ways.

2 Using Axioms to Reason About Goal Representations

The axiomatic approach starts with a set of simple properties and derives powerful conclusions about agents that satisfy them. For instance, we may assert that an ideal agent should exhibit transitive preference: if the agent prefers **A** to **B**, and **B** to **C**, it should prefer **A** to **C**. This property is one of four "VNM" axioms that can be used to conclude that "rational" agents are expected utility maximizers [24]. For an agent with multiple principals, we might further assert a Pareto axiom: if all principals prefer **A** to **B**, so too should the agent. This is used in Arrow's Impossibility Theorem [2] and Harsanyi's aggregation theorem [5], which together provide strong normative guidance: if we accept their axioms, our agents should have the ability to compare the utilities of different human principals.

Prior Work. A core focus of my PhD was to apply and extend the normatively satisfying results from decision theory and social choice to reinforcement learning based agents, where the existence of "rewards" that specify general purpose objectives was previously assumed without justification. My paper *Rethinking the Discount Factor* [14] was among the first to critically examine the traditional fixed discount, Markov reward paradigm in reinforcement learning, by showing that agents satisfying a Dynamic Consistency axiom [8] and the VNM axioms only need to satisfy a more expressive, *generalized* Bellman relation with state-action dependent discounting. This allows agents to represent a wider range of objectives than the standard fixed discount Bellman relation while still precluding inconsistencies, and suggests there are human preferences that traditional fixed discount reinforcement learning agents *cannot* represent. In my paper on *Consistent Aggregation of Objectives* [15], I combine the prior result with a Pareto condition to show that the aggregation of Markovian objectives with differing time preference must be non-Markovian. This tells us that general purpose agents should take into account past compromises when making present decisions, and prompts novel solutions to fundamental tensions in intertemporal choice.

Future Work. While the axiomatic approach has received significant attention in economics it remains underappreciated and underdeveloped in the field of AI. My goal is to continue to draw upon the significant economic literature, and to leverage new axioms and axiom combinations to gain insight into the structure of agent objectives. For example, I am currently studying the aggregation of subjective utilities that are may be based on different world models, and whether an objectivity axiom can be used in a multi-principal setup to address the objective social choice problem [21]? My hope is that this line links empirical human preference and "ideal" preferences, and makes progress on the critical question: *how can our reasoning about ideal objectives inform the real world agent design?* A critical limitation of current research is its focus on short horizon problems and short horizon supervision. To this end, I would like to continue to explore the questions of time preference and intertemporal choice prompted by my prior work [14, 15]: *how should we structure the rewards and time preferences of agents that operate over a long time horizon?*

3 Recognizing and Leveraging the Structure of Goals

To move beyond the representation theorems afforded by axioms and begin to reason about the substance of ideal goals, it is useful to recognize the structure inherent in the agent's external environment. Whereas axioms are asserted as desired properties of goals, structural assumptions are made with respect to an empirical quantity and informed by data. I argue that the most important structural assumptions are based on dependence and independence (e.g. statistical or causal), which unlock compositional generalization and allow for the application of axiomatically-justified aggregation theorems, as well as representational geometry (e.g. continuity and metric space assumptions), which allow us to reason about and generalize across spaces of goals and subgoals.

Prior Work. In my work on *Counterfactual Data Augmentation* [19, 20] we propose a powerful local generalization of causal modeling that applies even when global causal independence fails. It is shown, both theoretically and empirically, that leveraging local independence enables compositional generalization and allows reinforcement learning agents to generalize to unseen goals. A similar idea is used in my work on supervised reinforcement learning [12] to improve sample efficiency by stitching together conditionally independent subtrajectories. I've also explored how reinforcement learning agents can improve sample efficiency by exploiting geometric assumptions about extrinsic goal spaces in my work on *Maximum Entropy Gain Exploration for Long Horizon Multi-goal RL* [18] (proposed a state-of-the-art intrinsic goal setting algorithm for frontier exploration), prototype goal encodings [16] (uses a finer goal topology to solve coarse goals more efficiently), and neural networks that respect the triangle inequality [17] (models an agent's state and goal space as a quasi-metric space).

Future Work. Empirical human preferences are perhaps the best signal we have for studying ideal objectives. By investigating the causes of expressed preferences, we may gain a better understanding of the things humans truly care (or should care) about, and enable generalization across individuals and groups. For example, *how does the causal structure of the world, especially over longer time horizons, interact with human preferences?* Another interesting approach here is to leverage geometry to represent the manifold of human preferences and interpolate between diverse individuals, generalize across abstractions, and identify critical differences between groups.

4 Expressing Goals and Evaluating Agents Using Natural Language

With the advent of general-purpose language models (LMs) [11], AI agents can effectively process and respond to natural language inputs. This capability provides a powerful human-compatible interface for specifying and interpreting goals. With language, we might form a canon of interpretable principles and rules that characterize ideal agent behavior [23]. For example, Anthropic's Constitutional AI [3] applies a set of human-designed principles to steer LMs toward helpful and harmless behavior. While this has proven effective for shaping basic instruction following behaviors, natural language is inherently underspecified, which can lead to incomplete instructions, disagreement between human principals, and misunderstandings on the part of agents.

Prior Work. Our work *LLMs are Human-Level Prompt Engineers* [27] showed that effective meta instructions (e.g., system prompts) can be self-generated by LMs given sufficient coverage of target behaviors. In many applications, however, the range of possible inputs is so large that (a) target behaviors cannot be fully enumerated, and (b) any meta instruction would be too abstract or underspecified to provide proper guidance. In *Identifying the Risks of LM Agents* [13] ("ToolEmu") we use LMs to curate a set of test cases for LM tool use. We address the first issue by using an adversarial simulation and test case curation process that seeks out long-tailed failure modes of LM agents. We address the second by associating each test case with an "expected achievement" criteria that outlines desired, context-specific behavior. In my recent work on *Improving Context-Aware Preference Modeling* [22, 23] we investigate and improve the ability of LMs to evaluate behaviors given such context-specific criteria.

Future Work. The near-infinite range of potential LM behaviors poses significant challenges for traditional quantitative evaluations, as any fixed benchmark will fail to cover the full range of agent capabilities, limitations, and risks. In our ongoing work on *Report Cards* [25], we propose to address this by using LMs to generate qualitative evaluations of LM behavior by summarizing the most characteristic and semantically meaningful aspects of that behavior. This *ex-post*, human-interpretable approach to evaluation captures potentially unexpected nuances in model behaviors, and complements the adversarial approach to behavior elicitation used in ToolEmu [13]. While natural language is likely too underspecified to directly capture the full complexity of ideal general-purpose objectives, our work on context-aware preference modeling [22] finds that objectives with respect to specific criteria, or the objectives of a specific principal, can be captured with a high degree of accuracy ($\geq 90\%$ human agreement). This opens the door to my ongoing work on "PAgent", a modular, multi-principal simulation and evaluation framework where the agent's policy interacts with and impacts several diverse principals. This setup will allow us to safely study both the behavior of agents in complex, multi-principal settings, as well as the implications of different solution concepts (methods of aggregating the diverse preferences of the principals).

5 A Path Toward Alignment

My focus on the *normative* foundations of alignment makes my agenda unique relative to empirically-driven alignment plans, such as OpenAI's "iterative, empirical approach" [10] or Anthropic's "empiricism in AI safety" [1]. While I appreciate the immediate need to address the near-term empirical alignment problem that accompanies the deployment of our most capable models, I believe we can and should make simultaneous progress on the long-term technical alignment problem, which I argue is normative in nature. In pursuing the multi-faceted approach set out above, my aim is to contribute to a robust framework that lays the ethical and theoretical groundwork for deploying agents that pursue normatively justified objectives and advance humanity.

References

[1] Anthropic. Core views on AI safety. https://www.anthropic.com/index/core-views-on-ai-safety, 2023.

[2] Kenneth Joseph Arrow. Social choices and individual values. 1951.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Neural Information Processing Systems*, 2017.

[5] John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 1955.

[6] Hong Jun Jeon, Smitha Milli, and Anca D Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Neural Information Processing Systems*, 2020.

[7] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. Github, 2023.

[8] Mark J Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 1989.

[9] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

[10] OpenAI. Our approach to alignment research. https://openai.com/blog/our-approach-to-alignment-research, 2023.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. 2022.

[12] Keiran Paster, **Silviu Pitis**, Sheila McIlraith, and Jimmy Ba. Return augmentation gives supervised RL temporal compositionality. In *NeurIPS Deep RL Workshop*, 2022.

[13] Yangjun Ruan, Honghua Dong, Andrew Wang, **Silviu Pitis**, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024.

[14] **Silviu Pitis**. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *AAAI Conference on Artificial Intelligence*, 2019.

[15] **Silviu Pitis**. Consistent aggregation of objectives with diverse time preferences requires non-Markovian rewards. In *Neural Information Processing Systems*, 2023.

[16] **Silviu Pitis**, Harris Chan, and Jimmy Ba. ProtoGE: Prototype goal encodings for multi-goal reinforcement learning. In *Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2019.

[17] **Silviu Pitis**, Harris Chan, Kiarash Jamali, and Jimmy Ba. An inductive bias for distances: Neural nets that respect the triangle inequality. In *International Conference on Learning Representations*, 2020.

[18] **Silviu Pitis**, Harris Chan, Stephen Zhao, Bradly Stadie, and Jimmy Ba. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*, 2020.

[19] **Silviu Pitis**, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. In *Neural Information Processing Systems*, 2020.

[20] **Silviu Pitis**, Elliot Creager, Ajay Mandlekar, and Animesh Garg. MoCoDA: Model-based counterfactual data augmentation. In *Neural Information Processing Systems*, 2022.

[21] **Silviu Pitis** and Michael R. Zhang. Objective social choice: Using auxiliary information to improve voting outcomes. In *International Conference on Autonomous Agents and Multi-Agent Systems*, 2020.

[22] **Silviu Pitis**, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. Improving context-aware preference modeling for language models. In *Neural Information Processing Systems*, 2024.

[23] **Silviu Pitis**, Ziang Xiao, and Alessandro Sordoni. Canonical design for language agents using natural language reward models. In *Moral Philosophy and Moral Psychology Workshop at NeurIPS*, 2023.

[24] J Von Neumann and Oskar Morgenstern. Theory of games and economic behavior. 1953.

[25] Blair Yang, Fuyang Cui, Keiran Paster, Jimmy Ba, Pashootan Vaezipoor, **Silviu Pitis**, and Michael R. Zhang. Report cards: Qualitative evaluation of language models using natural language summaries. *Socially Responsible Language Modelling Research (SoLaR) at NeurIPS*, 2024.

[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[27] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, **Silviu Pitis**, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *International Conference on Learning Representations*, 2023.