Normative Agency Design

Research Statement
Silviu Pitis, November 2023

My ultimate research interest lies in the *normative design of general purpose artificial agency*. While empirical work provides valuable insights into agent behavior, the normative approach focuses on prescribing how agents should be designed, and is essential for developing AI systems that are ethically and socially aligned with human values. How *should* we design general purpose AI agents that benefit humanity? What objectives *should* agents pursue?

My research focuses on these questions in the context of designing reward-seeking, language-enabled agents that operate over a long time horizon, and applies tools from reinforcement learning, decision theory, natural language processing, causal inference, neural networks and representation learning. This statement first motivates the normative approach by examining the shortcomings of imitating human intelligence. Then it describes my past work and future plans.

1 Imitating Human Intelligence Is Not Enough

To define intelligence, we often use humans as a prototype. Animals exhibiting more human capabilities are said to be more intelligent. Similarly, we compare the task-specific performance of artificial agents (Deep Blue, AlphaGo, GPT, etc.) to human performance. For value alignment and reward learning, we appeal to human examples and preferences [6, 8, 11, 13]. We find this human-centric approach attractive because we *are* human—we understand ourselves and know what makes us intelligent. Or do we? For all our achievements, the brain's inner workings and our true motivations remain a mystery, we regularly exhibit undesirable behaviors and thought patterns [9], and our values and beliefs continue to rapidly evolve (e.g., just two centuries ago, slavery was commonly accepted). While we should aspire to create agents that possess human capabilities, any approach that is modeled *too closely* after the "empirical human" will suffer from the same shortcomings. To the extent we have irrational or anti-social humans, a strictly empirical approach may result in irrational or anti-social AIs (e.g., biases appearing in the data are propagated to trained models, and the chatbots Tay and WormGPT were deliberately misaligned).

2 Normative Agency Design

To separate our study of agency from the empirical human, we can imagine that there is some underlying truth or "ideal" that humanity is in the process of discovering (cf. Asimov's "Last Question" [3]). What purposes would an ideal agent pursue? Or put another way, what objectives would an ideal human (or society) give to our agents?

Due to our bounded rationality and limited experience, empirical humans provide only noisy, incomplete reflections of this ideal. However, through historical experience, ethics, philosophy, and mathematics, we have constructed certain generally accepted and internally consistent systems of thought. To move past the empirical human we may lean on such systems to inform agent design. This still involves human feedback, but large aspects of the problem are offloaded to more "objective" constructs, such as axiomatically-motivated aggregation mechanisms. I call this ***normative agency design*** because it specifies how agents *should* think and behave.

My research approaches normative design from three distinct angles: axiomatic, canonical and structural. I describe these below, in each case together with my work to date, related work, and future directions.

The Axiomatic Approach

The axiomatic approach starts with a set of simple, appealing 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 [25]. For an agent with multiple principals, we may consider axioms from social choice. For instance, the Pareto axiom—if all principals prefer **A** to **B**, so too should the agent—is used in both Arrow's Impossibility Theorem [2] and Harsanyi's aggregation theorem [7]. These results exemplify how combining simple axioms can provide strong normative guidance on agent design: if we accept their axioms, our agents should have a way to compare the utilities of different human principals.

Prior Work. A core focus of my PhD has been to extend and apply 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 has been assumed without justification. My paper *Rethinking the Discount Factor* [16]

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 [10] and the VNM axioms only need to satisfy the 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* [17], 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 realm of general purpose artificial agency. My goal is to leverage new axioms and axiom combinations to gain insight into the structure of rewards. For example, can subjective expected utility, together with novel abstraction and objectivity axioms, be applied in a multi-agent system to address the objective social choice problem [23]? 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?* Beyond ideal preferences, which may never be fully specified, axiomatic design may also be applied to decision making processes such as voting and policy making. *What properties ensure fairness, minimize market failures, and maximize overall welfare when true preferences are uncertain?* Finally, 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 [16, 17]: *how should we structure the rewards and time preferences of agents that operate over a long time horizon?*

The Canonical Approach

The canonical approach to normative design uses a set of natural language principles or criteria (a "canon") to inform behavior, reasoning, and judgments. As compared to axioms, these principles may be less self-evident, more heuristic, and lack a formal definition. However, they share the same normative flavor (the agent *should* do X) while also offering a path toward substantive alignment (axioms tend to capture consistency rather than substance). Similar to laws and morals, a particular principle may apply only in a limited set of circumstances, and will typically synthesize several specific experiences into a broadly applicable rule, although it may also take the form of a prototypical example (cf. case law). By accumulating a set of natural language principles, we may design agency that is interpretable, aligned, and transfers knowledge across modalities.

Prior Work. While the canonical approach finds roots in old-fashioned, knowledge-based AI, its application to machine learning-based agents was only made possible in the last few years by instruction-following Language Models (LMs) [13]. A simple way to construct a set of principles that governs AI is via human-engineering. For example, Anthropic's Constitutional AI [4] applies a set of human-designed principles to steer LMs toward helpful and harmless behavior. Manual prompt engineering can be burdensome, however, and our paper *LLMs are Human-Level Prompt Engineers* [26] shows that effective instructions for steering LM behavior can be self-generated by LMs—a significant step toward learning a rich canon from data. In *Identifying the Risks of LM Agents* [15] we use LMs to curate a set of test cases for LM tool use, each containing "expected achievement" criteria that outline desired LM behavior, which provides a proof-of-concept canon for LM tool use.

Future Work. My goal in this direction is to contribute to the design of a practical and socially accepted canon that is inspired by and functionally similar to human legal systems. To do this, we must first understand the best way to represent the canon and use it to steer LMs. How well LMs can understand and evaluate principles at different levels of specificity? How well can they evaluate the applicability of the principles themselves? I am actively exploring these questions [24] and considering other important directions, such as algorithms for principle learning and inference, and methods for principle generation, aggregation, improvement, and invalidation. An interesting hypothesis, parallel to the infamous reward hypothesis [5], is the *language reward hypothesis*: "all of what we mean by goals and purposes can be well expressed in natural language". Though I'm not sure the truth of this, I believe that significant progress can be made toward encoding agent goals into a set of generally accepted, human interpretable principles. If the principle following capabilities of LMs are sufficiently developed, this could open the door for governance that interfaces directly with agents and is self-enforced at the agent level.

The Structural Approach

The structural approach to normative design assumes a useful decomposition or property with respect to a process or thing, and uses it to make better predictions. How should the agent model the world? Whereas axioms are asserted as desired properties of an ideal, structural assumptions are made with respect to an empirical quantity and informed by data. A good example is Newton's Laws of Motion, which are extensively verified and strongly predictive for "classical" systems. The structural approach recognizes that the structure of the external environment, which can only be understood through interaction, is crucial to the optimal design of the agent. For AI, I argue that the most important structural assumptions are based on dependence and independence (e.g. statistical or causal), which unlock compositional generalization, as well as representational geometry (e.g. continuity and metric space assumptions), which allow us to reason about spaces of objects such as achievable goals.

Prior Work. In my work on *Counterfactual Data Augmentation* [21, 22] 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 tasks. A similar idea is used in my work on supervised reinforcement learning [14] 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* [20] (proposed a state-of-the-art intrinsic goal setting algorithm for frontier exploration), prototype goal encodings [18] (uses a finer goal topology to solve coarse goals more efficiently), and neural networks that respect the triangle inequality [19] (models an agents state and goal space as a quasi-metric space).

Future Work. Empirical human values are perhaps the best signal we have for ideal values. *Can the structure of empirical preferences inform our interpretation thereof, and allow us to better serve the underlying human values?* By studying 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, in order to interpolate between diverse individuals, explore the boundaries of human supervision, generalize across abstractions, and identify critical differences between groups. My goal here is to discover and leverage such empirical insights to inform the normative design of agent objectives.

3 A Path Toward Alignment

Each of these approaches—axiomatic, canonical, and structural—offers unique insights, but their true power lies in their complementary interplay. For example, axioms provide a formal foundation for aggregating diverse objectives, each of which may be expressed via a canonical principle. Meanwhile, the structural approach supports generalization between objectives and offers empirical backing for both the principles and axioms.

My focus on the *normative* foundations of alignment, particularly via the axiomatic approach, makes my agenda unique relative to empirically-driven alignment plans, such as OpenAI's "iterative, empirical approach" [12] 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. 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] Isaac Asimov. The last question. *Science Fiction and Philosophy*, 1956.

[4] 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.

[5] Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, 2023.

[6] 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.

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

[8] 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.

[9] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

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

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

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

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

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

[15] 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. *arXiv preprint arXiv:2309.15817*, 2023.

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

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

[18] **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.

[19] **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.

[20] **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.

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

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

[23] **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.

[24] **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.

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

[26] 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.