Normative Agency Design

Research Statement
Silviu Pitis, September 2023

I'm interested in the *normative design of general purpose agents*: how *should* we design agents that solve general tasks and contribute positively to society? What objectives *should* general purpose agents pursue?

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 (e.g., RLHF applied to LLMs), we appeal to human examples and preferences [10,13,18,19]. 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 [14], and our individual 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 to intelligence 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, such approach will result in irrational or anti-social AIs (e.g., Tay [33], WormGPT).

2 Normative Intelligence

To separate our study of agency from the empirical human, we can imagine that there is some underlying truth that humanity is in the process of discovering (cf. [5]). Due to our bounded rationality and limited experiences, empirical humans provide only noisy, incomplete reflections of this truth. However, through historical experience, philosophy, and mathematics, we have constructed certain generally accepted and internally consistent systems of thought, that provide some insight into the underlying truth. To move past the empirical human we may lean on these systems, and use them to inform the design of our agents. This still involves human feedback, but large aspects of the problem are offloaded to more "objective" constructs. I call this ***normative* agency design** because it specifies how agents *should* think and behave. I now describe three styles of normative design—axiomatic, canonical and structural—each with a bullet on related work.

The Axiomatic Approach. The axiomatic approach starts with a set of simple, appealing properties and uses them to derive 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 transitivity property is one of four "VNM" axioms that can be used to conclude that "rational" agents are expected utility maximizers [31]. For an agent with multiple principals, we should further consider axioms from social choice [4]. 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 [3] and Harsanyi's aggregation theorem [11]. These two results provide strong normative guidance on agent design: if we accept their axioms, our agents should have a way to compare the utilities of their human principals.

- In Rethinking the Discount Factor [21], I show that agents satisfying a Dynamic Consistency axiom [15,16] and the VNM axioms must satisfy a generalized Bellman relation with state-action dependent discounting [32]. 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 standard fixed discount reinforcement learning agents *cannot* represent. In my paper on Consistent Aggregation [22], 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. Interesting research from others relating to axiomatic design includes: [1,7,8,28,29,30].

The Canonical Approach. The canonical approach accumulates a set of principles to form a knowledge-base that is used to inform behavior, reasoning, and judgments, similar to the evolution of morality and law. Examples include the Ten Commandments, common law, the US Constitution, and Asimov's Laws of Robotics. As compared to axioms, these principles may be less self-evident, more heuristic, and lack a formal definition. While a particular principle may apply only in a limited set of circumstances, it will typically synthesize several specific experiences into a more broadly applicable rule. Research in this area may include proposing and analyzing principles, developing algorithms for principle learning and inference, and studying methods for principle collection, retrieval, application, transfer, analogical reasoning (common in legal arguments), aggregation, debate, improvement, and invalidation. By working with a natural language canon, and studying the ability of agents to follow and improve the canon, we may design agency that is interpretable, aligned, and transfers knowledge across modalities.

- The application of a natural language canon to AI was made possible by instruction-tuned LLMs [19]. The most obvious way to construct a set of principles that governs AI is via human-engineering [2, 12]. For example, Constitutional AI [6] applies a set of human-designed principles to steer LLMs toward more helpful and harmless behavior. Manual prompt engineering can be difficult, and our Automatic Prompt Engineer paper [34] shows that effective instructions can be self-generated by LLMs, which is the first step is learning a rich canon from data. This is something I am quite interested in actively working toward. A natural hypothesis, parallel to the infamous reward hypothesis [7], is the *language reward hypothesis*: "all of what we mean by goals and purposes can be well expressed in natural language".

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. Whereas axioms are asserted as desired properties of a thing, structural assumptions are made with respect to an empirical quantity and typically arrived to via an inductive process. A good example is Newton's Laws of Motion, which are extensively verified and strongly predictive for "classical" systems. For AI, I would argue that the most important structural assumptions are based on dependence and independence (e.g. statistical or causal), as well as representational geometry (e.g. continuity and metric space assumptions).

- Of the three styles of normative design, the structural approach is *by far* the broadest in terms of existing work (e.g. [17]). In terms of my own work, I have proposed to exploit local causal independence for counterfactual data augmentation and shown that this can allow agents to generalize to unseen tasks [26, 27]. A similar idea is used by my work on data augmentation for reward-conditioned reinforcement learning [20] to improve sample efficiency of learning. I've also had a few papers that exploit geometric assumptions, including my works on Maximum Entropy Gain Exploration [25] (proposes intrinsic goal setting for frontier exploration), Prototype Goal Encodings [23] (uses specific goals to achieve general ones), and Neural Metrics [24] (models the goal space as a quasi-metric space).

3 A Path Toward Superhuman Alignment

These three styles of normative design—axiomatic, canonical and structural—are all closely related, and in some ways overlapping. They each work with smaller, simpler pieces to reason about the whole. If the decomposition and aggregation mechanism are valid, then human-centric supervision on the pieces is sufficient to supervise the whole, creating a pathway to superhuman intelligence and scalable alignment [9].

Of course, once the superhuman comes (e.g. language, calculus, industrial automation, the Internet), the hope is that it fast becomes human, and through accumulated experience, we can start to supervise it directly, and use it as a component toward the next level of "superhuman".

References

[1] D. Abel, W. Dabney, A. Harutyunyan, M. K. Ho, M. Littman, D. Precup, and S. Singh. On the expressivity of markov reward. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[2] M. Anderljung, J. Barnhart, J. Leung, A. Korinek, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

[3] K. J. Arrow. Social choices and individual values. 1951.

[4] K. J. Arrow, A. Sen, and K. Suzumura. *Handbook of social choice and welfare*, volume 2. Elsevier, 2010.

[5] I. Asimov. The last question. *Science Fiction and Philosophy*, pages 279–289, 1956.

[6] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[7] M. Bowling, J. D. Martin, D. Abel, and W. Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023.

[8] H. Cao, S. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

[9] P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

[10] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[11] J. C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955.

[12] L. Ho, J. Barnhart, R. Trager, Y. Bengio, M. Brundage, A. Carnegie, R. Chowdhury, A. Dafoe, G. Hadfield, M. Levi, et al. International institutions for advanced ai. *arXiv preprint arXiv:2307.04699*, 2023.

[13] H. J. Jeon, S. Milli, and A. D. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *arXiv preprint arXiv:2002.04833*, 2020.

[14] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[15] T. C. Koopmans. Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society*, pages 287–309, 1960.

[16] M. J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(4):1622–1668, 1989.

[17] A. Mohan, A. Zhang, and M. Lindauer. Structure in reinforcement learning: A survey and open problems. *arXiv preprint arXiv:2306.16021*, 2023.

[18] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[19] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[20] K. Paster, S. Pitis, S. McIlraith, and J. Ba. Return augmentation gives supervised rl temporal compositionality. *NeurIPS 2022 Deep RL Workshop*, 2022.

[21] S. Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the 33nd AAAI Conference on Artificial Intelligence*. AAAI Press, 2019.

[22] S. Pitis. Multi-objective agency requires non-markovian rewards. *Under Review*, 2023.

[23] S. Pitis, H. Chan, and J. Ba. ProtoGE: Prototype goal encodings for multi-goal reinforcement learning. In *The 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2019.

[24] S. Pitis, H. Chan, K. Jamali, and J. Ba. An inductive bias for distances: Neural nets that respect the triangle inequality. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*, 2020.

[25] S. Pitis, H. Chan, S. Zhao, B. Stadie, and J. Ba. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *Proceedings of the Thirty-seventh International Conference on Machine Learning (ICML)*, 2020.

[26] S. Pitis, E. Creager, and A. Garg. Counterfactual data augmentation using locally factored dynamics. In *Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[27] S. Pitis, E. Creager, A. Mandlekar, and A. Garg. Mocoda: Model-based counterfactual data augmentation. 2022.

[28] M. Shakerinava and S. Ravanbakhsh. Utility theory for sequential decision making. In *International Conference on Machine Learning*, pages 19616–19625. PMLR, 2022.

[29] J. Skalse and A. Abate. On the limitations of markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In *Uncertainty in Artificial Intelligence*, pages 1974–1984. PMLR, 2023.

[30] J. M. V. Skalse, M. Farrugia-Roberts, S. Russell, and A. Gleave. Invariance in policy optimisation and partial identifiability in reward learning. 2023.

[31] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior. 1953.

[32] M. White. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3742–3750. JMLR. org, 2017.

[33] M. J. Wolf, K. W. Miller, and F. S. Grodzinsky. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12, 2017.

[34] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2023.