Normative Agency Design

Research Statement
Silviu Pitis, November 2020

I'm interested in the *normative design of general purpose artificial agency*: how *should* we design AIs that solve general tasks and contribute positively to society? The first part of this document motivates normative agency design and summarizes my past work. The second part outlines my current work and related ideas.

I Normative Agency Design and Past Work

Imitating Human Intelligence Is Not Enough. To define intelligence, we often use the human 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, DQN, AlphaGo, etc.) to human performance. For value alignment and reward learning, we typically consider the human as the "gold standard" [5,13,17,22]. We find this human-centric approach attractive because we *are* human—we know how we think and what makes us intelligent. **Or do we?** For all our achievements, the inner workings of the brain and our true motivations remain a mystery, we regularly exhibit undesirable behaviors and thought patterns [18], and our individual values and beliefs continue to 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 [44]).

Normative Intelligence. To separate our study of agency from the empirical human, we can describe the qualities we seek in an "ideal agent", and design our agents so as to satisfy them. I call this ***normative* agency design** because it specifies how agents *should* think and behave. I now describe two styles of normative design—axiomatic and capability/constraint—together with a bullet summarizing my past work.

The Axiomatic Approach. The axiomatic approach to normative design starts with a set of basic, universally appealing axioms and uses them to derive powerful conclusions about agents that satisfy the axioms. The axioms can be understood as defining "rationality" and are usually simple, intuitive statements. For instance, we may think 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 axiom is one of four VNM axioms that can be used to conclude that "rational" agents are expected utility maximizers [23,42], which is fundamental to our use of value functions in reinforcement learning (RL) [37]. For an agent with multiple principals (e.g., any pro-social agent), 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 [15]. 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.

- The work I am most proud of is my work on discounting [28], in which I show that agents satisfying a Dynamic Consistency axiom [19,21], the VNM axioms and a few other axioms must satisfy a *generalized* Bellman relation with state-action dependent discounting. This provides a normative justification for White's unifying RL task formalism [43], and allows agents to represent a wider range of objectives than the standard Bellman relation, while still precluding inconsistencies. If we accept my axiomatization, it suggests there are human preferences that fixed discount agents *cannot* represent (i.e., *reward learning is not enough—we must also learn time preference*). I plan to empirically verify this in future work.

The Capability/Constraint Approach. The capability/constraint approach to normative design asks "what capabilities should agents possess?" and similarly, "what constraints should we place on agents?", and proceeds to construct agents with these capabilities and constraints. Work in this category has a more empirical flavor than axiomatic work, because capabilities are often defined in less formal terms, and much like written law, some interpretation is required. Nevertheless, by cataloging desired capabilities and constraints, we can start to form a more clear picture of the ideal agent. For instance, the fact that a general purpose agent must explore, communicate and adapt through time, has motivated my focus on the RL setting [40].

- I take this approach in my works on Source Traces (grants agents a model of potential causes) [27], Maximum Entropy Gain Exploration (MEGA) (grants agents the ability to systematically explore an environment) [31], Prototype Goal Encodings (ProtoGE) (grants agents the ability to pursue goals from multiple goal spaces) [29], and Counterfactual Data Augmentation (CoDA) (grants agents the ability to generate counterfactuals by mixing independent subsamples of observed transitions) [32] and Neural Metrics (architecturally constrains an agent's value function to satisfy the triangle inequality) [30]. While I like this research and think it is valuable, I find it somewhat less satisfying than the axiomatic approach. The solutions are mathematically motivated, but they exemplify only one, not necessarily best solution.

II Research Topics I'm Interested In (November 2020)

One of my long run goals is to obtain a clear normative account of intelligence *before* we develop a general purpose, potentially superhuman, AI. My immediate interests, reward aggregation and flexible multi-goal specification, primarily relate to principled abstraction—a critical piece of the larger puzzle. As it is likely our agents, regardless of compute, will possess but *bounded rationality* [39], normative design demands principled approximation. Yet perhaps the most common modern approach is to throw a neural network at the problem and approximate *everything*—entire observations, entire states, even entire trajectories [6,12,14]. This produces representations and behaviors that are entangled in subtle, often obscure ways with the data distribution, optimization process, and hyperparameters [1, 9, 16]. Even more so than the empirical human, the resulting agents are black boxes with unclear motivations, possibly inconsistent beliefs [20], and poor generalizability [7]. I don't yet have a clear picture of what the correct approach is, but I believe it should involve some form of compositional, context-conditioned reasoning over basic abstractions. This motivates my interest in multi-goal RL [26, 29, 30, 31, 32, 34, 38], which allows agents to plan over a range of goals, typically defined as abstractions (sets) over the state space. Below, I outline some of my current ideas.

Decomposing Individual Intelligence. We typically characterize individual choice as the result of a monolithic set of preferences. Consider, however, the corporation or national government. We can treat them as individual agents, but we can also view their actions as the result of a negotiation between several human principals (I used both lenses to analyze corporate and national policy from a law and economics angle in law school [24, 25]). This perspective also applies to individual action, as a negotiation between competing objectives [36], and so we can apply the axioms and tools of social choice to analyze individual choice. At present, I'm using this approach to work toward an universal basis for reward functions. My current theoretical work shows that under the assumptions of my work on discounting [28] (including the VNM and Dynamic Consistency axioms), the Pareto axiom, and the presence of varying time preference [8], it is *impossible* to consistently combine reward functions without forgoing the Markov property (cf. [45]). This suggests that to faithfully represent compound objectives, we must treat rewards as non-Markovian [41]. Can this analysis be applied to design agents that can successfully navigate conflicting objectives and principals?

Aggregating Imperfect Principals. The prior paragraph considers an agent with multiple fully observed, cardinal principals. But how should an agent integrate noisy feedback signals? [35]. My recent project on Objective Social Choice [33] formalizes the noisy aggregation problem in the bandits setting by framing voting as an estimation problem. We found that auxiliary information can be used to improve outcomes and concluded that we should not assume Anonymity—which treats all principals equally—as an axiom. Can this be extended to the sequential setting, where current approaches (e.g., [5]) are Anonymous? Relatedly, what should we do when signals are a mix of ordinal and cardinal values?

Axiomatizing Abstraction. A normative approach to abstraction could be of tremendous value. For instance, my labmate and I recently designed a neural network architecture that is guaranteed to respect the triangle inequality [30]. The intent was to use these networks to represent asymmetric metrics in multi-goal RL, thereby providing our agents with consistency guarantees on their value estimates. We discovered, however, that the underlying idea—representing distances in RL using an asymmetric metric—can be applied naively only in the non-abstracted case, where goals are completely specified states. In the abstracted case, the triangle inequality may fail and we need a more nuanced approach, which we are now working toward. If we had normative guidance, e.g., on how to represent and/or compute distances between abstractions, it might simplify our problem significantly. Can we come up with an axiomatic approach to abstraction?

Enabling Dynamic Abstraction. One of our most remarkable human capabilities is that we can view the same underlying state from a range of diverse perspectives. In a legal trial, for instance, the lawyers for each side attempt to persuade the judge and jury that their characterization of the facts is the "better" one. I believe this ability to reason in terms of multiple abstractions is one of the most important capabilities for an agent to possess: it is necessary for empathy, compromise, and communication with others whose basic assumptions differ. As the right low-dimensional abstraction can also greatly simplify a variety of tasks, including exploration [30, 31], the ability to represent abstractions dynamically [10, 11], as well as translate between abstractions [29], would be invaluable. How can we grant our agents the ability to reason about the world from multiple competing perspectives? And what makes a good abstraction?

Constraining Intent. The usual "global reward" approach to RL is *intent-free*, or at best, *intent implicit*. An intent-free agent that does something wrong is difficult to debug: did it fail because the goal was misspecified or because the learning algorithm failed? Multi-goal RL grants agents explicit intent, but this comes with risks: a “curious” block stacking agent seeks to drop the block in novel ways [31]. Borrowing an idea from criminal law, we may constrain bad intent (*mens rea*); but how exactly should this be done?

Procedural normativity. When faced with hard substantive problems, we might borrow another idea from the legal system: provide the necessary procedure (e.g., due process) so that ultimate outcomes are deemed fair, even if they are considered “wrong” by some (or many!). We can take a normative approach to this and impose axiomatic requirements on our procedure. For instance, we use this approach to analyze ordinal voting rules; even though no rule can satisfy our normative desiderata (recall Arrow’s Theorem [3]), we can nevertheless require rules to satisfy fairness axioms like Anonymity, Condorcet Consistency and Strategy-Proofness. When allowing humans to shape the objectives of AIs, are there any procedural properties—separate from substantive rationality constraints—that we can impose to prevent negative externalities?

Stacking Six Blocks From Scratch. A nice benchmark problem for sparse-reward multi-goal reinforcement learning is block stacking. Stacking 2 blocks from scratch is too hard for a randomly exploring Hindsight Experience Replay agent [2]. Our MEGA agent [31] is the first to solve this task from scratch, without a hand-crafted curriculum or human demonstrations. In fact, by adding a safety constraint (see “Constraining Intent” above), we can even stack 3 blocks from scratch. But I have not yet been successful in stacking 6 blocks. Is this possible by combining techniques like MEGA, CoDA [32] and safety constraints?

References

[1] M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, et al. What matters in on-policy reinforcement learning? A large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020.

[2] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5048–5058, 2017.

[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] 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.

[6] J. D. Co-Reyes, Y. Liu, A. Gupta, B. Eysenbach, P. Abbeel, and S. Levine. Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings. *arXiv preprint arXiv:1806.02813*, 2018.

[7] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pages 1282–1289. PMLR, 2019.

[8] S. Frederick, G. Loewenstein, and T. O’Donoghue. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.

[9] J. Fu, A. Kumar, M. Soh, and S. Levine. Diagnosing bottlenecks in deep q-learning algorithms. *arXiv preprint arXiv:1902.10250*, 2019.

[10] M. Garnelo, D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. J. Rezende, and S. Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018.

[11] J. Gordon, J. Bronskill, M. Bauer, S. Nowozin, and R. E. Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

[12] D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pages 2450–2462, 2018.

[13] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[14] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

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

[16] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. *arXiv preprint arXiv:1709.06560*, 2017.

[17] 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.

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

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

[20] T. Lu, D. Schuurmans, and C. Boutilier. Non-delusional q-learning and value-iteration. In *Advances in neural information processing systems*, pages 9949–9959, 2018.

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

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

[23] S. Pitis. Examining expected utility theory from descriptive and prescriptive perspectives. *Independent undergraduate research; advised by Professor Archishman Chakraborty*, 2010.

[24] S. Pitis. Designing optimal takeover defenses. *Independent law school research; advised by Professor Holger Spamann*, 2013.

[25] S. Pitis. Punitive damages in international trade. *Independent law school research; advised by Professor Mark Wu*, 2013.

[26] S. Pitis. Reasoning for reinforcement learning. In *Hierarchical Reinforcement Learning Workshop at the 31st Conference on Neural Information Processing Systems (HRL@NIPS)*, 2017.

[27] S. Pitis. Source Traces for temporal difference learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.* AAAI Press, 2018.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] S. Pitis and M. R. Zhang. Objective social choice with non-i.i.d. votes. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[34] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

[35] A. D. Procaccia. How is voting theory really useful in multiagent systems? *available online, URL: http://www. cs. cmu. edu/arielpro/papers/vote4mas. pdf (DOA: 15.01. 2013).*

[36] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[37] S. Russell. Rationality and intelligence: A brief update. In *Fundamental issues of artificial intelligence*, pages 7–28. Springer, 2016.

[38] T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320, 2015.

[39] H. A. Simon. Models of man; social and rational. 1957.

[40] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

[41] S. Thiébaux, C. Gretton, J. Slaney, D. Price, and F. Kabanza. Decision-theoretic planning with non-markovian rewards. *Journal of Artificial Intelligence Research*, 25:17–74, 2006.

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

[43] 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.

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

[45] S. Zuber. The aggregation of preferences: can we ignore the past? *Theory and decision*, 70(3):367–384, 2011.