On Serving Two Masters

Directing Critical Technical Practice towards Human-Compatibility in AI

McKane Andrus
mckaneandrus@berkeley.edu
UC Berkeley Department of Computer Science

KEYWORDS

human compatible AI; critical technical practice; critical HCI; discourse analysis; commitment analysis; shared control

ACM Reference Format:
McKane Andrus. 2019. On Serving Two Masters: Directing Critical Technical Practice towards Human-Compatibility in AI. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306618.3314325

INTRODUCTION

In this project I have worked towards a method for critical, socially-aligned research in Artificial Intelligence by merging the analysis of conceptual commitments in technical work, discourse analysis, and critical technical practice. While the goal of critical technical practice as proposed by [1] is to overcome technical impasses, I explore an alternative use case - ensuring that technical research is aligned with social values. In the design of AI systems, we generally start with a technical formulation of a problem and then attempt to build a system that addresses that problem. Critical technical practice tells us that this technical formulation is always founded upon the discipline's core discourse and ontology, and that difficulty in solving a technical problem might just result from inconsistencies and faults in those core attributes. What I hope to show with this project is that, even when a technical problem seems solvable, critical technical practice can and should be used to ensure the human-compatibility of the technical research.

I specifically focus on a recent extension of Human-Compatible Artificial Intelligence, Negotiable Reinforcement Learning (NRL) [2], which aims to adjudicate the desires and beliefs of multiple individuals through the use of optimal planning/control. Following the method outlined in [7], I conduct a conceptual commitment analysis centered around NRL's definitions of "Negotiation" and "Negotiability" to uncover the constraints those definitions impose. Then, I engage in a slightly deeper dive into the nascent field of Human Compatible AI to explore how the discourse frames both the problems NRL is supposed to solve and NRL's contributions to the field. Finally, in the spirit of critical technical practice, I developed and implemented an algorithm and system for NRL. Having carried out this three-pronged critical analysis, I commit the project to alternative definitions of "Negotiation" drawn from the external fields of Computer Supported Cooperative Work, Negotiation and Group Decisions, and Social Psychology. From these commitments, I developed alternative algorithms and explore their relationships with the constraints set by the initial "Negotiation" definition. I then implemented these algorithms in a functional prototype as a way of demonstrating how interdisciplinary, critical alignments can inform technical design and be used to rigorously test the conceptual commitments on which the design rests. A next step for this project is to test whether the performance of these algorithms is indeed more aligned with human needs and values than the baseline. This task is exceedingly difficult, however, in that the NRL problem statement is not yet applicable to any real world use-cases. I believe the most important next step for strengthening my proposed methodology is to apply it to a more pervasive use-case of AI, as this would allow for the use of tools from Value-Sensitive Design and Participatory Design to gauge the methodology's success [4, 6].

Negotiable Reinforcement Learning

Negotiable Reinforcement Learning is defined by [2] as a formalization of the cooperative-control problem, in which there is a single robot attempting to satisfy multiple owners. This formalization begins with the assumption that each owner has a belief-model of the robot. These beliefs can be directly modeled as a POMDP, where beliefs over the effects of the robot's actions are transition probabilities, and where beliefs over the robot's observations are just observation probabilities. On top of this, each owner has a personalized reward function defined over their POMDP. For our purposes, we consider a more tractable formalization that assumes full observability of the world state. We are thus left with an MDP for each owner.

$$\forall_i \; MDP_i = (S, A, T_i, R_i, \gamma,) \tag{1}$$

In order to merge these into a single formalization, we add a hidden parameter θ to the set of MDPs. We also augment the transition and reward functions in the following manner:

$$\begin{gathered} MDPs = \forall_i \;\; (S, A, T, R, \gamma, \theta_i) \\ T(s, a, s', \theta_i) = P(s, a, s'|\theta_i) = T_i(s, a, s') \\ R(s, a, \theta_i) = R_i(s, a) \end{gathered} \tag{2}$$

We now have a Mixed-Observability MDP, or more specifically a Hidden-Parameter MDP (HiPMDP) [3], to represent a composite of the owners' models. By keeping beliefs over the θ_is, we maintain weights over each owner. In this setting, transitions act as observations of θ, biasing the weights towards the owners with the models that best predict the observed history. Belief updates are then:

$$b(\theta_i|s, a, s') = \frac{P(s, a, s'|\theta_i)b(\theta_i)}{\sum_j P(s, a, s'|\theta_j)b(\theta_j)} \tag{3}$$

The core theorem of [2] states that the optimal conditional plan for this formalization is Pareto optimal in expectation.

AIES '19, January 27–28, 2019, Honolulu, HI, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6324-2/19/01.
https://doi.org/10.1145/3306618.3314325

COMMITMENT ANALYSIS

For much of this part of the project I turn to a singular commitment that NRL is motivated by and arguably founded upon, negotiability as preference aggregation. As described in previous sections, NRL is used to maximize subjective expected utility, a quantification of preferences, suggesting that the recombination of these quantifications is the sole step in arriving at a desirable solution. Turning to two fields that have more deeply considered what it means for something to be negotiable, Computer Supported Cooperative Work (CSCW) and Group Decisions and Negotiation (GDN), I point to a number of limitations imposed by the simplified definition of negotiation and discuss how they impact the human-compatibility of the proposed model. For the sake of brevity, I only list the limitations here: Statically Optimal Solution Models, Outcome over Process, Limited Interaction between Users and Agent, Limited Interactions amongst Users, and Undifferentiated Treatment Between Contexts.

DISCOURSE ANALYSIS

Within the inchoate field of Human Compatible AI (closely related to, if not equivalent to, the field of AI Safety), the imagined context for the technologies and techniques being developed is one where AI is significantly more powerful than it is today. In this context, AI necessarily brings an increase in personal, organizational, and governmental capability. As such, there seems to be an immense need for AI that allows for shared control to prevent abuse of power or even all out AI arms races. The broader ideology that this hints at, however, is one that centers technical solutions over more sociopolitical approaches. Stemming from this line of inquiry, we explore the impacts of the following limitations: Solutionism, Responsibility Detachment, and the Obfuscation of Empirical Optimality.

TECHNICAL IMPLEMENTATION ANALYSIS

Finally, by developing an algorithm that efficiently solves the NRL POMDP formalization and subsequently building out environments to test it on, I open NRL to one final line of critique. Again, for the sake of brevity, I cannot go into much detail on the process of implementation or generation of critiques. The main concern that I develop, however, is that the solutions the NRL formalization effectively only settle bets. Any significant difference in beliefs encourages the agent to learn which user is most correct and simply favor that user from there on out. Future work will explore this concern in more detail.

ALTERNATIVE DESIGNS

From CSCW, GDN, and social psychology, we find that two core components of negotiation omitted from negotiation as preference aggregation are conflict identification and alternative solution generation and critique. Addressing each of these components individually, I developed two separate general algorithms.

For the first, we need a formulation of NRL that treats conflict identification as its central mechanism for facilitating negotiation. As such, we require a mechanism that, given the beliefs and preferences of the parties, can find a course of action that effectively calls attention to the beliefs or preferences that give rise to the most differentiation in desired plans. Our general approach to an NRL agent of this type relies upon information-gain based exploration in a similar fashion to [5]. Then, during the process of its operation, the agent returns information it has gathered in some interpretable way, allowing for reprogramming if the users so desire. Once the users believe they are in sufficient agreement on the state of the world, they can inform the agent, which then implements a more standard Multi-Objective Reinforcement Learning algorithm, as in [8], instead of one that incorporates differences of beliefs.

For the second, we need a formulation of NRL that centers alternative solution generation and critique as its mechanism for enabling negotiability. Given that we are working with agents meant to carry out complex plans in the virtual and/or real worlds, the combinatorial number of possible plans makes this a daunting task. Furthermore, a key component of solution proposal and critique is for users to understand, at least to some degree, how their input is considered in generating solutions. Thus, a first step in the generic approach is to show each user some possible outcomes if the agent were to only consider that user's own beliefs and preferences. The next step would be to show some possible outcomes in the case where all parties' beliefs and preferences are considered. After these steps, it might be possible for users to propose, attack, and defend both sets of beliefs and preferences and the solutions that could stem therefrom, but it is likely that specific use-cases would require highly specialized interfaces and algorithmic interpretability to ensure that these tasks are feasible for the intended users.

ALGORITHMIC IMPLEMENTATIONS

I implemented the original and two additional designs in a toy-environment where they could be tested with real humans. Given both the immense difficulty of belief- and preference-elicitation and the intractability of POMDPs in settings with high belief-space dimensionality, testing these designs in anything but a toy-environment is largely unfeasible. While I still need to gather more data for the toy-environment to make any final claims, I worry that the domain of NRL is just too abstract to meaningfully engage different publics. For this reason I'd like to apply my methodology to more grounded research domains, such as the treatment of 'fairness' in Machine Learning.

REFERENCES

[1] Philip E Agre. 1997. *Computation and human experience.* Cambridge University Press.

[2] Andrew Critch and Stuart Russell. 2017. Servant of Many Masters: Shifting priorities in Pareto-optimal sequential decision-making. *arXiv preprint arXiv:1711.00363* (2017).

[3] Finale Doshi-Velez and George Konidaris. 2016. Hidden parameter Markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *IJCAI: proceedings of the conference*, Vol. 2016. NIH Public Access, 1432.

[4] Batya Friedman, David G Hendry, Alan Borning, et al. 2017. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction* 11, 2 (2017), 63–125.

[5] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems.* 1109–1117.

[6] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[7] Merel Elisabeth Noorman. 2009. *Mind the gap: a critique of human/technology analogies in artificial agents discourse.* Maastricht University.

[8] Kristof Van Moffaert and Ann Nowé. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research* 15, 1 (2014), 3483–3512.