Characterizing the Limits of Autonomous Systems

Extended Abstract

Junzhe Zhang
Purdue University
West Lafayette, Indiana
zhang745@purdue.edu

Eias Bareinboim
Purdue University
West Lafayette, Indiana
eb@purdue.edu

ABSTRACT

This note answers two central question in the intersection of decision-making and causal inference – when human input is needed and, if so, how it should be incorporated into an AI system. We introduce the counterfactual agent who proactively considers human input in its decision-making process. We prove that a counterfactual agent dominates the standard autonomous agent who does not consider any human input (i.e., the experimental agent) in terms of performance. These results suggest a trade-off between autonomy and optimality – while the full autonomy is often preferred, using human input could potentially improve the efficiency of the system. We further characterize under what conditions experimental and counterfactual agents can reach the same level of performance, which elicits the settings where full autonomy can be achieved.

CCS CONCEPTS

• **Mathematics of computing → Causal networks**; • **Computing methodologies → Causal reasoning and diagnostics**;

KEYWORDS

Knowledge Representation and Reasoning; Human-robot/agent interaction

ACM Reference Format:
Junzhe Zhang and Eias Bareinboim. 2018. Characterizing the Limits of Autonomous Systems. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

One of the primary goals of reinforcement learning (RL) is to build an autonomous system where the agent operates independently in the environment performing complex tasks. Today, autonomous systems have been deployed in a wide range of settings, including autonomous driving [16], game-playing [13], and energy conservation [12]. In reality, many complex systems involve humans interacting and intervening in the environment. For most standard RL approaches (including Markov decision processes (MDPs) [11] and partially observable MDPs (POMDPs) [14, 15]), however, the human component is assumed to be oblivious [18]. Semi-autonomous systems (SAS) explicitly model the human's behavior, allowing the agent to consider its own as well as humans' capabilities proactively [18]. Using the language of structural causal models (SCMs) [2, 7, 10], Bareinboim et al. [1] designed a semi-autonomous system in the Multi-Armed bandit (MAB) settings, which proactively takes into account humans' decisions using counterfactual inference (and expanded in [5, 6, 8, 17]). In practice, however, the MAB is a rather simplistic model of the environment that is not applicable in the general sequential decision-making tasks where actions affect not only the immediate rewards but also future states of the agent.

In this work, we model the sequential decision-making environments using the language of SCMs. We define the *counterfactual agent* using the human input in its decision process (i.e., the semi-autonomous systems). We compare the performance of the counterfactual agent and the standard autonomous systems, which do not consider any human input, called the *experimental agent*. One perhaps surprising result of our theory is that the counterfactual agent could dominate its experimental counterpart, *even when* the human operator performs poorly, potentially worse than random guessing. These results allude to a trade-off between autonomy and optimality: while full autonomy is often preferable, a semi-autonomous approach that leverages human's capability could potentially achieve higher performance.

2 PLANNING WITH HUMAN INPUT

We start the discussion with an example involving the sequential planning of patients' treatment, but, obviously, this example can be extended to any decision-making setting. In this case, a physician treats each patient who visits the hospital regularly to maintain her long-term health condition. The physician measures the patient's corticosteroid level at the t-th visit, $S^{(t)} = s^{(t)}$. She then decides a treatment $X^{(t)} = x'^{(t)}$, and then measures an overall health score $Y^{(t)} = y^{(t)}$. In reality, the patient's health score $Y^{(t)}$ is also affected by a pair of confounders $U^{(t)} = \{M^{(t)}, E^{(t)}\}$, where $M^{(t)} = m^{(t)}$ stands for patient's psychological status and $E^{(t)} = e^{(t)}$ stands for her socioeconomic status. $M^{(t)}, E^{(t)}$ cannot be accessed by anyone except the physician due to privacy concerns, thus considered as unobserved in our setting. The patient's long-term health condition is measured by the γ-discounted cumulative reward. The physician decides a treatment $X^{(t)} = x'^{(t)}$ following a policy function $x'^{(t)} = \pi_h(s^{(t)}, m^{(t)}, e^{(t)})$. We model this sequential environment using a structural causal model M [10, Ch. 7] described in Fig. 1 (for $t = 1, 2$), which we refer to as the MDPUC model.

To maximize the patient's long-term health, the hospital's administration aims to automate the decision procedure and deploy an experimental agent A_{exp}. The agent A_{exp} decides an action $x^{(t)}$ on the basis of its visited states $s^{([1,t])}$ and actions $x^{([1,t-1])}$, regardless of the physician's decisions $x'^{([1,t])}$.[1] We solve for this

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

[1] We use $v^{([i,j])}$ to denote a sequence $\langle v^{(i)}, v^{(i+1)}, \ldots, v^{(j)} \rangle$ (empty if $j < i$).

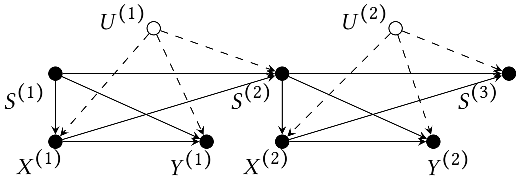

Figure 1: Causal diagram for MDPUC where $U^{(t)}$ affects only $\{X^{(t)}, Y^{(t)}, S^{(t+1)}\}$.

system using both the standard MDP and POMDP planning algorithms and label the resulting policies as *exp* and *exp2*, respectively. We also include two baseline policies for comparison: (1) the physician's current policy (called *human*), and (2) a policy picking the treatment at random (called *random*). Fig. 2 shows the cumulative reward of these policies. The physician (*human*) performs worse than random guessing (*random*). Perhaps surprisingly, the performance of *exp* and *exp2* coincide with the *random* policy, indicating that the autonomous approach A_{exp} is unable to learn a reasonable policy, which is expected to be better than chance.

3 COUNTERFACTUAL AGENTS

Recall that the experimental agent A_{exp} always ignores the physician's decision $x'^{(t)}$, which may contain valuable information about the hidden state. We could improve its decision-making procedure by proactively considering the physician's preference, namely,

Definition 3.1 (Counterfactual Agent). A counterfactual agent A_{ctf} is defined as a policy $\Pi_{ctf} = \pi^{([1,t])}$ where $\pi^{(t)}$ is a function deciding the action $X^{(t)} \leftarrow \pi(s^{([1,t])}, x'^{([1,t])}, x^{([1,t-1])})$.

At state $s^{(t)}$, the counterfactual agent takes the human's decision $x'(t)$ as its intended action, infer the outcome $y^{(t)}, s^{(t+1)}$ *had it taken* a different action $x^{(t)}$ (counterfactually), and finally make a decision. We could obtain the optimal policy of the counterfactual agent using the standard POMDP planning methods [3, 4, 9] by using the human's decision $x'^{(t)}$ as a partial observation of the hidden state. However, solving for POMDPs is computationally hard in practice. By exploiting the structure of MDPUCs using counterfactual logic, an efficient solution can be obtained. Let the counterfactual variable Y_x be the solution of Y in the submodel where the functions associated with variables X are replaced with constants $X = x$ [10, Ch. 7.1]. We show that the Markov property [11] holds for a counterfactual agent in MDPUC environments.

Theorem 3.2 (Counterfactual Markov Property). *For an MDPUC model, the following equalities hold for $t \geq 1$,*

$$P\big(s^{(t+1)}_{x([1,t])}, x'^{(t+1)}_{x([1,t])} \mid s^{([1,t])}_{x([1,t-1])}, x'^{([1,t])}_{x([1,t-1])}\big)$$
$$= P\big(s^{(t+1)}_{x(t)}, x'^{(t+1)}_{x(t)} \mid s^{(t)}, x'^{(t)}\big),$$
$$E\big[y^{(t)}_{x([1,t])} \mid s^{([1,t])}_{x([1,t-1])}, x'^{([1,t])}_{x([1,t-1])}\big] = E\big[y^{(t)}_{x(t)} \mid s^{(t)}, x'^{(t)}\big],$$

where $s^{([1,t])}_{x([1,t-1])} = \{s^{(t)}_{x([1,t-1])}\}_{t\geq 1}, x'^{([1,t])}_{x([1,t-1])} = \{x'^{(t)}_{x([1,t-1])}\}_{t\geq 1}.$

Thm. 3.2 says that in MDPUCs, the observed state $s^{(t)}$ and the human decision $x'^{(t)}$ (at time t) perfectly summarize the history of a counterfactual agent A_{ctf}. We are then allowed to treat the human decision $x'^{(t)}$ as if it were part of the observed state, and solve for A_{ctf} using the standard MDP algorithms [11].

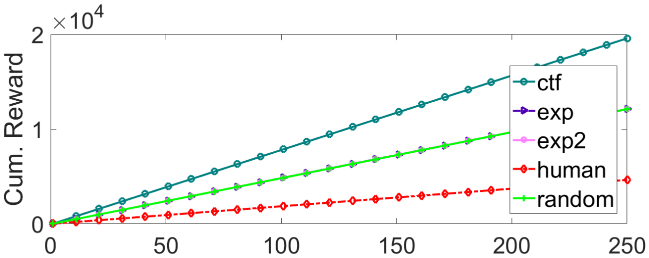

Figure 2: Simulations comparing performance of experimental (A_{exp}) and counterfactual (A_{ctf}) agents in offline planning. X-axis represents the total episodes.

We compute the optimal policy of A_{ctf} in the MDPUC model and label it as *ctf*, which is shown in Fig. 2. The results reveal that the counterfactual agent A_{ctf}, leveraging the human's capabilities, consistently outperforms the experimental agent A_{exp}, notably, even when the human performs worse than random guessing.

3.1 Autonomy vs. Optimality

The results discussed so far suggest a trade-off between optimality and autonomy when designing RL systems – while full autonomy is preferable, the agent could potentially achieve better performance by leveraging the human's capabilities, even if the information affecting the human decision has an adverse effect on its performance. Let the optimal discounted expected cumulative rewards for A_{exp} be denoted by V^*_{exp} and for A_{ctf} by V^*_{ctf}. One can formally show the relationship between the cumulative rewards, V^*_{exp} and V^*_{ctf}.

Theorem 3.3. *For an MDPUC, $V^*_{exp} \leq V^*_{ctf}$. If the human decision is affected only by observed variables ($U^{(t)} \not\to X^{(t)}$), $V^*_{exp} = V^*_{ctf}$.*

This proposition confirms the intuition that a counterfactual agent A_{ctf} dominates any experimental agent A_{exp}, which follows from having access to additional information. The result further confirms that whenever the human operator does not have access to (or is influenced by) any piece of information hidden to the agent, one could replace the human with an autonomous agent without sacrificing the performance, i.e., full autonomy can, at least in principle, be achieved.

4 CONCLUSION

We introduced counterfactual agents and showed both theoretically and empirically that they dominate standard autonomous agents (experimental) in terms of quality of the solution obtained. Given that in real-world settings human decision-makers are almost invariably influenced by unobserved confounders, our findings suggest that human input should generally be considered, which perhaps surprisingly, it's true even when humans are worse than chance. Our characterization delineates a formal boundary for the performance achieved by semiautonomous and fully-autonomous systems in a wide variety of natural and artificial decision-making scenarios.

REFERENCES

[1] Elias Bareinboim, Andrew Forney, and Judea Pearl. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*. 1342–1350.

[2] E. Bareinboim and J. Pearl. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113 (2016), 7345–7352. Issue 27.

[3] Anthony Cassandra, Michael L Littman, and Nevin L Zhang. 1997. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann, 54–61.

[4] Hsien-Te Cheng. 1988. *Algorithms for partially observable Markov decision processes.* Ph.D. Dissertation. University of British Columbia.

[5] Nicolás Della Penna, Mark D Reid, and David Balduzzi. 2016. Compliance-Aware Bandits. *arXiv preprint arXiv:1602.02852* (2016).

[6] A. Forney, J. Pearl, and E. Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. 1156–1164.

[7] J.Y. Halpern. 1998. Axiomatizing Causal Reasoning. In *Uncertainty in Artificial Intelligence*, G.F. Cooper and S. Moral (Eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

[8] Finnian Lattimore, Tor Lattimore, and Mark D Reid. 2016. Causal Bandits: Learning Good Interventions via Causal Inference. In *Advances in Neural Information Processing Systems*. 1181–1189.

[9] Michael L Littman. 1994. The witness algorithm: Solving partially observable Markov decision processes. (1994).

[10] J. Pearl. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York.

[11] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming.* Wiley & Sons.

[12] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, Richard Dazeley, et al. 2013. A Survey of Multi-Objective Sequential Decision-Making. *J. Artif. Intell. Res.(JAIR)* 48 (2013), 67–113.

[13] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[14] Satinder P Singh, Tommi S Jaakkola, and Michael I Jordan. 1994. Learning Without State-Estimation in Partially Observable Markovian Decision Processes.. In *ICML*. 284–292.

[15] Richard D Smallwood and Edward J Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* 21, 5 (1973), 1071–1088.

[16] Chris Urmson, Joshua Anhalt, Drew Bagnell, Christopher Baker, Robert Bittner, MN Clark, John Dolan, Dave Duggins, Tugrul Galatali, Chris Geyer, et al. 2008. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics* 25, 8 (2008), 425–466.

[17] Junzhe Zhang and Elias Bareinboim. 2017. Transfer Learning in Multi-Armed Bandits: A Causal Approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

[18] Shlomo Zilberstein. 2015. Building Strong Semi-autonomous Systems. In *Proceedings of the Twenty-Ninth AAAI Conference*. AAAI Press, 4088–4092.