Research Statement On AI Safety

Usman Anwar

Arguably, AI safety is *the* problem of the century [10]. Developing solutions for AI safety is not only critical to ensure the survival and prosperity of humanity in the long term but also to ensure that public trust in AI systems does not get eroded in the short term [1, 11]. Even though the current AI technologies are quite limited, there is an already growing concern among the larger public about their harmful effects [2, 12, 14, 23]. In nuclear energy, we already have an example of a high-impact technology that has remained under-utilized due to adverse shift in public opinion caused by large-scale disasters due to poor safety protocols [4, 25]. I am a firm believer in the power of AI technologies to have a transformative effect on human society, so, ensuring that AI does not suffer a similar fate to nuclear energy and remains a safe and reliable technology is the primary motivation for me to study safe AI. Luckily, the AI safety problem [1, 6, 11, 16, 30] also possesses great intellectual appeal for me. It blends challenges from artificial intelligence, engineering, philosophy, and economics in a way that I find fascinating. I have formal training in electrical engineering and machine learning and deep interests in philosophy and economics; so, an area of research at the intersection of these fields is an enticing prospect for me. I expect to make technical contributions towards developing practical methods to make AI safer. In the following sections, I briefly discuss three research themes relevant to AI safety that I will explore during my Ph.D. research.

1 Inverse Constrained Reinforcement Learning

Imposing constraints on an agent i.e. specifying what the agent *must* not do is a natural way of ensuring that the agent acts safely. Human society uses constraints and penalties in form of laws, scripted rules, and social norms to ensure that members of the society act safely [19]. Constraints are also used in machine learning to prohibit known bad behaviors. For example, fairness constraints [38] are used in fair machine learning to prevent machine learning models from disregarding low data classes and safety constraints are used in safe reinforcement learning to promote safety [27, 3, 5, 9]. However, manual specification of constraints is quite hard [22, 32, 37] and therefore it is prudent to think of other ways in which constraints could be specified.

Inverse Constrained Reinforcement Learning (ICRL) is a framework for learning safety constraints from the demonstrations of safe behavior by a human or another agent. Initially proposed for tabular settings with deterministic transition dynamics, my prior works extend this framework to high dimensional continuous settings [22] and stochastic dynamics [26]. More importantly, my work shows that learned safety constraints are quite robust and transfer extremely well to new agents which may have different dynamics, embodiments, and even reward functions. This is quite significant and indicates that for any new domain, constraints only need to be learned once. This is in contrast to imitation learning and inverse reinforcement learning (IRL) where the learned reward models often fail to transfer adequately [34].

I further hypothesize that a shining property of ICRL is that it is a viable way to provide scalable supervision [6]. Specifically, because the safety constraints are compositional[1], we can learn constraint sets over small domains and then aggregate them to get a constraint set for a larger domain. For a practical example, consider a house with k rooms and an AI-controlled robot that has to act in all the rooms. While IL and IRL will require demonstrations that explore the full house jointly, using ICRL one can learn constraints separately for each room and then aggregate them together to learn a constraint set for the complete house. Recursively, constraints from multiple homes can then be aggregated as well in a similar way and so on. Empirical verification of this hypothesis and understanding the pros and cons of the proposed approach above is an obvious direction of research for me.

In parallel, I intend to work on enhancing the capabilities of the ICRL framework, so that it can learn non-Markovian constraints, work in environments with partial observability and use off-policy data to be data efficient. Further, because the current ICRL methods only learn constraints, any new agent when trained on these constraints is likely to first violate them and then slowly learn to respect them. This is undesirable and I hope that in the future I can provide a solution that mitigates this issue. Finally, despite the similarity in learning from demonstrations, ICRL and IRL are different frameworks targeted at learning different things; I am excited about uncovering scenarios in which both can be used simultaneously and may complement each other as well.

2 Impact Regularizers

Any AI agent that may pose an existential threat to humanity is bound to have a large impact on the world. Thus, limiting the impact of an AI agent is one way to make it safer. Based on this intuition, [8] introduced the concept of “low impact AI" which dictates that an AI system must always be designed with the auxiliary objective that it must make as little changes to the world state as possible. This effect can be achieved by using impact regularizers which penalize the actions with disproportionate impact. Despite the intuitive appeal, the practical application of this idea suffers because of the lack of scalable approaches to measure the impact of an AI system on the world. As direct measurement of impact is generally very difficult [8], the community has mostly focused on defining useful proxies that reliably correlate with the impact of an AI system. A popular proxy is attainable utility preservation [17, 20] which forces the agent to perform well under auxiliary reward functions, specified by a human, in addition to its own primary reward function or objective [17]. Recently, it has been observed that in some cases using a randomly generated reward function as the sole auxiliary objective also works well [35]. This is quite surprising and warrants further investigation. A useful diagnosis of the unreasonable effectiveness of a random reward function as an auxiliary objective can shed light on the necessary properties of a good auxiliary objective and may help in the design of novel auxiliary objectives. In addition, I plan on developing methods to automate the specification of auxiliary reward functions using insights from automatic goal generation methods [13].

[1] Compositionality means that if a domain A has constraint set C_A and domain B has constraint set C_B, then the superdomain formed by joining domains A and B will have the constraint set $C_A \cup C_B$.

3 Distribution Generalization

Another important challenge for ensuring AI safety is distribution generalization[2] [6]. Without distribution generalization, there is no guarantee that AI agents acting in dynamic real-world environments will preserve their safe behavior over the long term even if initially verified to be safe. Despite development of novel machine learning methods with enhanced distribution generalization [7, 18, 36] properties in recent years, there are concerns that in practice these methods do not provide any significant gains over traditional machine learning methods in terms of distribution generalization [15, 24, 28]. An alternative line of research that I believe has been under-researched is using novel forms of supervision. For example, explanations have been found to be an effective method for improving robustness and generalization of machine learning models [29, 31, 33]. My goal is to find novel and innovative forms of supervisions that may help in achieving distribution generalization as well as develop a theoretical understanding of the essential ingredients, in terms of supervision, needed for generalization. A good starting point for this will be understanding what forms of supervision enable learning causally correct models [21].

4 Contributions To The Field

The unsafe behavior from AI agents can emanate from two major sources: lack of clear specifications to AI agent about the desired objective and failure of the AI agent to learn a behavior that robustly optimizes the desired objective. I hope to make contributions on both of these fronts. I expect my work on ICRL to provide a scalable approach towards learning safety constraints from data. This will enable the AI agents to know about the potentially unsafe behaviors and avoid them. I am particularly excited about the prospects of aggregating learned constraints from multiple domains; this can potentially scale to AI systems working at a scale where collecting direct demonstrations from humans is infeasible thus making it impossible to use IRL. With impact regularizers, my goal is to transform them to the level where they could be used in plug-n-play fashion, not much different than regularizers in optimization. This will enable practitioners to trivially combine them with their own objectives and achieve safer behavior from AI agents. Finally, with my work on distribution generalization, I desire to pursue a direction that is relatively orthogonal to the approach of the community. Considering the limited real success that the community has had with the more popular approaches, I believe my contributions here can prove pivotal if successful and lay the foundation of an alternative approach towards distribution generalization based on providing richer supervisions.

On the whole, because the planned contributions are not tied to any specific algorithms or applications, I expect them to remain applicable even for larger scale AI systems. In particular, I am quite hopeful that ICRL framework and my associated works will be directly useful in promoting safe behaviours of strong AI systems that may act at much larger scales compared to current AI systems and may be helpful for building safe Artificial General Intelligence (AGI).

[2] Distribution generalization is also referred to as out-of-domain generalization

References

[1] Benefits & risks of artificial intelligence. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/, 2018. Accessed: 20 October, 2021.

[2] A. Abid, M. Farooqi, and J. Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

[3] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, 2017.

[4] J. F. Ahearne. Nuclear power after chernobyl. *Science*, 236(4802):673–679, 1987.

[5] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

[6] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety, 2016. arXiv:1606.06565.

[7] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[8] S. Armstrong and B. Levinstein. Low impact artificial intelligences. *arXiv preprint arXiv:1705.10720*, 2017.

[9] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

[10] A. Cotra. Why ai alignment could be hard with modern deep learning. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/, 2021. Accessed: 20 October, 2021.

[11] A. Critch and D. Krueger. Ai research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*, 2020.

[12] C. Dougherty. Google photos mistakenly labels black people 'gorillas'. *The New York Times*, 2015. Accessed: 20 October, 2021.

[13] C. Florensa, D. Held, X. Geng, and P. Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528. PMLR, 2018.

[14] M. Foundation. Youtube regrets a crowdsourced investigation into youtube's recommendation algorithm. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf, 2021. Accessed: 20 October, 2021.

[15] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[16] H. Karnofsky. Potential risks from advanced artificial intelligence: the philanthropic opportunity. https://www.openphilanthropy.org/blog/

potential-risks-advanced-artificial-intelligence-philanthropic-opportunity#Some_Open-Phil-specific_considerations, 2016.

[17] V. Krakovna, L. Orseau, R. Ngo, M. Martic, and S. Legg. Avoiding side effects by considering future tasks. *arXiv preprint arXiv:2010.07877*, 2020.

[18] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 2021.

[19] J. D. Lewandowski. Capitalising sociability: Rethinking the theory of social capital. In *Assessing social capital: Concept, policy and practice*, volume 14, pages 14–28. Cambridge Scholars Publishing in association with GSE Research, 2012.

[20] D. Lindner, K. Matoba, and A. Meulemans. Challenges for using impact regularizers to avoid negative side effects. *arXiv preprint arXiv:2101.12509*, 2021.

[21] D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

[22] S. Malik, U. Anwar, A. Aghasi, and A. Ahmed. Inverse constrained reinforcement learning. In *International Conference on Machine Learning*, pages 7390–7399. PMLR, 2021.

[23] G. McDonald. Danger, danger! 10 alarming examples of ai gone wild. https://www.infoworld.com/article/3184205/danger-danger-10-alarming-examples-of-ai-gone-wild.html#slide4, 2017. Accessed: 20 October, 2021.

[24] V. Nagarajan, A. Andreassen, and B. Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fSTD6NFIW_b.

[25] O. of Nuclear Energy. Advantages and challenges of nuclear energy. https://www.energy.gov/ne/articles/advantages-and-challenges-nuclear-energy, 2021. Accessed: 20 October, 2021.

[26] D. Papadimitriou, U. Anwar, and D. Brown. Bayesian inverse constrained reinforcement learning. In *NeurIPS 2021 Workshop on Safe and Robust Control of Uncertain Systems*, 2021.

[27] A. Ray, J. Achiam, and D. Amodei. Benchmarking safe exploration in deep reinforcement learning, 2019. https://cdn.openai.com/safexp-short.pdf.

[28] E. Rosenfeld, P. K. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

[29] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[30] S. Saisubramanian, S. Zilberstein, and E. Kamar. Avoiding negative side effects due to incomplete knowledge of ai systems. *arXiv preprint arXiv:2008.12146*, 2020.

[31] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[32] D. R. Scobee and S. S. Sastry. Maximum likelihood constraint inference for inverse reinforcement learning. In *International Conference on Learning Representations*, 2020.

[33] W. Stammer, P. Schramowski, and K. Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3619–3629, 2021.

[34] S. Toyer, R. Shah, A. Critch, and S. Russell. The magical benchmark for robust imitation. *arXiv preprint arXiv:2011.00401*, 2020.

[35] A. M. Turner, D. Hadfield-Menell, and P. Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–391, 2020.

[36] E. Vinitsky, Y. Du, K. Parvate, K. Jang, P. Abbeel, and A. Bayen. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.

[37] T.-Y. Yang, M. Hu, Y. Chow, P. J. Ramadge, and K. Narasimhan. Safe reinforcement learning with natural language constraints. *arXiv preprint arXiv:2010.05150*, 2020.

[38] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 2019.