Tractable Probabilistic Models for Ethical AI

Vaishak Belle(✉)

University of Edinburgh & Alan Turing Institute, London, UK
vbelle@ed.ac.uk

Abstract. Among the many ethical dimensions that arise in the use of ML technology, three stand out as immediate and profound: enabling the interpretability of the underlying decision boundary, addressing the potential for learned algorithms to become biased against certain groups, and capturing blame and responsibility for a system's outcomes. In this talk, we advocate for a research program that seeks to bridge tractable (probabilistic) models for knowledge acquisition with rich models of autonomous agency that draw on philosophical notions of beliefs, intentions, causes and effects.

1 Motivation

Machine learning (ML) techniques have become pervasive across a range of different applications, and are now widely used in areas as disparate as recidivism prediction, consumer credit-risk analysis, and insurance pricing [8,25]. Likewise, in the physical world, machine learning models are critical components in autonomous agents such as robotic surgeons and self-driving cars. Among the many societal/ethical dimensions that arise in the use of ML technology in such applications, three stand out as immediate and profound. First, to increase trust and accommodate human insight, interpretability of the underlying decision boundary is essential. Second, there is the potential for learned algorithms to become biased against certain groups, which needs to be addressed. Third, in so much that the decisions of ML models impact society, both virtually (e.g., denying a loan) and physically (e.g., accidentally driving into a pedestrian), the enabling of blame and responsibility is a significant challenge.

Many definitions have been proposed in the literature for such ethical considerations [2,19], but there is considerable debate about whether a formal notion is appropriate at all, given the rich social contexts that occur in human-machine interactions. Valid arguments are also made about the challenges of model building and deployment [10,11]: everything from data collection to ascribing responsibility when technology goes awry can demonstrate and amplify abuse of power

This article is a written version of the keynote to be given at *27th International Conference on Conceptual Structures*, September 2022, Münster, Germany. A preliminary version of this work was also presented at the *Critical Perspectives on Artificial Intelligence Ethics Conference* in Edinburgh, UK, 2020. The author was supported by a Royal Society University Research Fellowship.

T. Braun et al. (Eds.): ICCS 2022, LNAI 13403, pp. 1–6, 2022.
https://doi.org/10.1007/978-3-031-16663-1_1

and privilege. Such issues are deeply intertwined with legal and regulatory problems [15,32].

Be that as it may, what steps can be taken to enable ethical decision-making a reality in AI systems? Human-in-the-loop systems are arguably required given the aforementioned debate [24,34], but such loops still need to interface with an automated system of considerable sophistication that in the very least reasons about the possible set of actions. In particular, simply delegating responsibility of critical decisions to humans in an ad hoc fashion can be problematic. Often critical actions can be hard to identify immediately and it is only the ramification of those actions that raise alarm, in which case it might be too late for the human to fix. Moreover, understanding the model's rationale is a challenge in itself, as represented by the burgeoning field of explainable artificial intelligence [4,14,29]. So a careful delineation is needed as to which parts are automated, which parts are delegated to humans, which parts can be obtained from humans a-priori (i.e., so-called *knowledge-enhanced machine learning* [9]), but also how systems can be made to reason about their environment so that they are able to capture and deliberate on their choices, however limiting their awareness of the world might be. In the very least, the latter capacity offers an additional layer of protection, control and explanation before delegating, as the systems can point out which beliefs and observations led to their actions.

2 Two-Pronged Approach

In that regard, our view is that a two-pronged approach is needed in the least. On the one hand, we have to draw on philosophical notions and look to formalise them, as attempted by the knowledge representation community. Indeed this community has looked to capture beliefs, desires, intentions, and causality in service of formal notions that provide an idealised perspective on epistemology grounded in, say, a putative robot's mental state [5,16,21,22]. But the topic of knowledge acquisition, i.e., how the relevant propositions can be acquired from data is largely left open. Moreover, the topic of reasoning, i.e., of computing truths of acquired knowledge is a long-standing challenge owing to the intractability of propositional reasoning and the undecidability of first-order logic, and many higher-order logics.

On the other hand, although machine learning systems do successfully address acquisition from data, mainstream methods focus on atomic classification tasks, and not the kind of complex reasoning over physical and mental deliberation that humans are adept in. (There are exceptions from robotics and reinforcement learning, of course, but these all attempt some form of mental state modeling [1], and in the very least, reasoning about possible worlds [31].) Moreover, issues about robustness in the presence of approximate computations remain.

In this talk, we advocate for a research program that seeks to apply tractable (probabilistic) models to problems in fair machine learning and automated reasoning of moral principles. Such models are compilation targets for propositional

and finite-domain relational logic, and so can represent certain types of knowledge representation languages. They can also be learned from data. We report on a few preliminary results [7,12,17,23,26–28,33,35]. Firstly, we discuss results on studying causality-related properties in such models, and extracting counterfactual explanations from them. On the topic of fairness, it is shown that the approach enables an effective technique for determining the statistical relationships between protected attributes and other training variables. This could then be applied as a pre-processing step for computing fair models. On the topic of moral responsibility, it is shown how models of moral scenarios and blameworthiness can be extracted and learnt automatically from data as well as how judgements be computed effectively. In both themes, the learning of the model can be conditioned on expert knowledge allowing us to represent and reason about the domain of interest in a principled fashion.

3 Closing Remarks

We conclude with key observations about the interplay between tractability, learning and knowledge representation in the context of ethical decision-making. Among other things, we observe that the tractable model paradigm is in its early years, at least as far capturing a broad range of knowledge representation languages is concerned, and moreover, there is altogether less emphasis on mental modeling and agency. (First-order expressiveness is yet another dimension for allowing richness in specifications, as are proposals with an explicit causal theory such as [30]). In contrast, readers may want to consult discussions in [23,24] on knowledge representation approaches where a more comprehensive model of the environment and its actors is considered, but where knowledge acquisition and learning are used in careful, limited ways.

Analogously, we observe that although many expressive languages [6,18] are known to compile to tractable models, this is purely from the viewpoint of reasoning, or more precisely, probabilistic query computation. What is likely needed is a set of strategies for reversing this pipeline: from a learned tractable model, we need to be able to infer high-level representations. In the absence of general strategies of that sort, the more modest proposal is perhaps to interleave declarative knowledge for high-level patterns but allow low-level patterns to be learnt, which then are altogether compiled for tractable inference.

Overall, the discussed results can be seen occupying positions on a spectrum: the fairness result simply provides a way to accomplish de-biasing, but does not engage with a specification of the users or the environment in any concrete way. Thus, it is closer to mainstream fairness literature. The moral reasoning result is richer in that sense, as it explicitly accounts for actors and their actions in the environment. However, it does not explicitly infer how these actions and effects might have come about – these might be acquired via learning, for example – nor does it reason about what role these actions play amongst multiple actors in the environment. Thus, clearly, in the long run, richer formal systems are needed, which might account for sequential actions [3] and multiple agents [20].

However, this reverts the position back to the issues of tractability and knowledge acquisition not being addressed in such proposals. So, the question is this: can we find ways to appeal to tractable probabilistic models (or other structures with analogous properties) with such rich formal systems? As mentioned, it is known that certain probabilistic logical theories can be reduced to such structures, so perhaps gentle extensions to those theories might suggest ways to integrate causal epistemic models and tractable learning.

Beyond that technical front, much work remains to be done, of course, in terms of delineating automated decision-making from delegation and notions of accountability [13]. It is also worth remaking that computational solutions of the sort discussed in the previous section do make strong assumptions about the environment in which the learning and acting happens. In a general setting, even data collection can amplify positions of privilege, and moreover, there are multiple opportunities for failure and misspecification [10,11]. Orchestrating a framework where this kind of information and knowledge can be communicated back to the automated system is not at all obvious, and is an open challenge. In that regard, the two-pronged approach is not advocated as a solution to such broader problems, and indeed, it is unclear whether abstract models can imbibe cultural and sociopolitical contexts in a straightforward manner. However, it at least allows us to specify norms for human-machine interaction, provide goals and situations to achieve, model the machine's beliefs, and allow the machine to entertain models of the user's knowledge. Ultimately, the hope is that the expressiveness argued for offers additional protection, control and explanation during the deployment of complex systems with machine learning components.

References

1. Albrecht, S.V., Stone, P.: Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif. Intell. **258**, 66–95 (2018)
2. Allen, C., Smit, I., Wallach, W.: Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. **7**(3), 149–155 (2005)
3. Batusov, V., Soutchanski, M.: Situation calculus semantics for actual causality. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
4. Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. arXiv preprint arXiv:2009.11698 (2020)
5. Brachman, R.J., Levesque, H.J., Reiter, R.: Knowledge Representation. MIT Press (1992)
6. Broeck, G.V.D., Thon, I., Otterlo, M.V., Raedt, L.D.: DTProbLog: a decision-theoretic probabilistic prolog. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, pp. 1217–1222. AAAI Press (2010)
7. Choi, Y., Dang, M., Broeck, G.V.D.: Group fairness by probabilistic modeling with latent fair decisions. arXiv preprint arXiv:2009.09031 (2020)
8. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
9. Cozman, F.G., Munhoz, H.N.: Some thoughts on knowledge-enhanced machine learning. Int. J. Approximate Reasoning **136**, 308–324 (2021)
10. Crawford, K.: The Atlas of AI. Yale University Press, New Haven (2021)

11. Crawford, K.: The hidden costs of AI. New Sci. **249**(3327), 46–49 (2021)
12. Darwiche, A.: Causal inference using tractable circuits. arXiv preprint arXiv:2202.02891 (2022)
13. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6
14. Doshi-Velez, F., et al.: Accountability of AI under the law: the role of explanation. arXiv preprint arXiv:1711.01134 (2017)
15. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J. Ethics **21**(4), 403–418 (2017)
16. Fagin, R., Moses, Y., Halpern, J.Y., Vardi, M.Y.: Reasoning About Knowledge. MIT Press, Cambridge (2003)
17. Farnadi, G., Babaki, B., Getoor, L.: Fairness in relational domains. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 108–114 (2018)
18. Fierens, D., Van den Broeck, G., Thon, I., Gutmann, B., De Raedt, L.: Inference in probabilistic logic programs using weighted CNF's. In: Proceedings of UAI, pp. 211–220 (2011)
19. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (IM) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016)
20. Ghaderi, H., Levesque, H., Lespérance, Y.: Towards a logical theory of coordination and joint ability. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 1–3 (2007)
21. Halpern, J.Y.: Actual Causality. MIT Press, Cambridge (2016)
22. Halpern, J.Y.: Reasoning About Uncertainty. MIT Press, Cambridge (2017)
23. Hammond, L., Belle, V.: Learning tractable probabilistic models for moral responsibility and blame. Data Min. Knowl. Disc. **35**(2), 621–659 (2021). https://doi.org/10.1007/s10618-020-00726-4
24. Kambhampati, S.: Challenges of human-aware AI systems. AI Mag. **41**(3), 3–17 (2020)
25. Khandani, A., Kim, J., Lo, A.: Consumer credit-risk models via machine-learning algorithms. J. Bank. Finan. **34**, 2767–2787 (2010)
26. Papantonis, I., Belle, V.: Interventions and counterfactuals in tractable probabilistic models. In: NeurIPS Workshop on Knowledge Representation & Reasoning Meets Machine Learning (2019)
27. Papantonis, I., Belle, V.: Closed-form results for prior constraints in sum-product networks. Frontiers Artif. Intell. **4**, 644062 (2021)
28. Papantonis, I., Belle, V.: Principled diverse counterfactuals in multilinear models. arXiv preprint arXiv:2201.06467 (2022)
29. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
30. Salimi, B., Parikh, H., Kayali, M., Getoor, L., Roy, S., Suciu, D.: Causal relational learning. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pp. 241–256 (2020)
31. Sardina, S., De Giacomo, G., Lespérance, Y., Levesque, H.J.: On the limits of planning over belief states under strict uncertainty. In: KR vol. 6, pp. 463–471 (2006)
32. Stilgoe, J.: Machine learning, social learning and the governance of self-driving cars. Soc. Stud. Sci. **48**(1), 25–56 (2018)
33. Varley, M., Belle, V.: Fairness in machine learning with tractable models. Knowl. Based Syst. **215**, 106715 (2021)

34. Zanzotto, F.M.: Human-in-the-loop artificial intelligence. J. Artif. Intell. Res. **64**, 243–252 (2019)
35. Zečević, M., Dhami, D., Karanam, A., Natarajan, S., Kersting, K.: Interventional sum-product networks: causal inference with tractable probabilistic models. In: Advances in Neural Information Processing Systems, vol. 34 (2021)