Knowledge Representation and Acquisition for Ethical AI: Challenges and Opportunities

Ethics and Information Technology (2023) 25:22
https://doi.org/10.1007/s10676-023-09692-z

ORIGINAL PAPER

Knowledge representation and acquisition for ethical AI: challenges and opportunities

Vaishak Belle[1]

Abstract
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. Likewise, in the physical world, ML models are critical components in autonomous agents such as robotic surgeons and self-driving cars. Among the many ethical dimensions that arise in the use of ML technology in such applications, analyzing morally permissible actions is both immediate and profound. For example, there is the potential for learned algorithms to become biased against certain groups. More generally, in so much that the decisions of ML models impact society, both virtually (e.g., denying a loan) and physically (e.g., driving into a pedestrian), notions of accountability, blame and responsibility need to be carefully considered. In this article, we advocate for a two-pronged approach ethical decision-making enabled using rich models of autonomous agency: on the one hand, we need to draw on philosophical notions of such as beliefs, causes, effects and intentions, and look to formalise them, as attempted by the knowledge representation community, but on the other, from a computational perspective, such theories need to also address the problems of tractable reasoning and (probabilistic) knowledge acquisition. As a concrete instance of this tradeoff, we report on a few preliminary results that apply (propositional) tractable probabilistic models to problems in fair ML and automated reasoning of moral principles. Such models are compilation targets for certain types of knowledge representation languages, and can effectively reason in service some computational tasks. They can also be learned from data. Concretely, current evidence suggests that they are attractive structures for jointly addressing three fundamental challenges: reasoning about possible worlds + tractable computation + knowledge acquisition. Thus, these seems like a good starting point for modelling reasoning robots as part of the larger ecosystem where accountability and responsibility is understood more broadly.

Keywords Machine learnin · Ethics in AI · Knowledge representation · Socio-technical systems

Introduction

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 (Chouldechova, 2017; Khandani et al., 2010). Likewise, in the physical world, ML models are critical components in autonomous agents such as robotic surgeons and self-driving cars. Among the many ethical dimensions that arise in the use of ML technology in such applications, analyzing morally permissible actions is both immediate and profound. For example, there is the potential for learned algorithms to become biased against certain groups. More generally, in so much that the decisions of ML models impact society, both virtually (e.g., denying a loan) and physically (e.g., driving into a pedestrian), notions of accountability, blame and responsibility need to be carefully considered.

Many definitions have been proposed in the literature for such ethical considerations (Friedler et al., 2016; Allen et al., 2005), 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 about model building and deployment (Crawford, 2021a, b): everything from data collection to denouncing responsibility when technology goes awry can demonstrate and amplify abuse

✉ Vaishak Belle
vbelle@ed.ac.uk

[1] University of Edinburgh & Alan Turing Institute, Edinburgh, UK

Published online: 11 March 2023

of power and privilege. Such issues are deeply intertwined with legal and regulatory problems (Etzioni & Etzioni, 2017; Stilgoe, 2018).

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 (Zanzotto, 2019; Kambhampati, 2020; Crootof et al., 2022), 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 (Rudin, 2019; Doshi-Velez et al., 2017; Belle & Papantonis, 2020). 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* (Cozman & Munhoz, 2021)), 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.

Main thesis

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, time, space, abstraction and causality in service of formal notions that provide an idealised perspective on epistemology grounded in, say, a putative robot's mental state (Brachman et al., 1992; Fagin et al., 2003; Halpern, 2016, 2017; Beckers & Halpern, 2019; Belle, 2020a; Reiter, 2001). 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 ML 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 (Albrecht & Stone, 2018), and in the very least, reasoning about possible worlds (Sardina et al., 2006.)) Moreover, issues about robustness in the presence of approximate computations remain.

As a concrete instance of this tradeoff, we report on a few preliminary results that apply (propositional) tractable probabilistic models (TPMs) to problems in fair ML and automated reasoning of moral principles. Such models are compilation targets for certain types of knowledge representation languages, and can effectively reason in service of some computational tasks. More recently, they have been shown an alternative scheme to encode joint distributions, permitting many probabilistic computations (conditional probabilities, marginals, expectations) to be efficient. Consequently, they are now being learned directly from data. In particular, current evidence suggests that they are attractive structures for jointly addressing three fundamental challenges[1]:

- reasoning about possible worlds (as required by logics of knowledge, intentions and norms) +
- tractable computation (as required for real-time behavior and/or scalability) +
- knowledge acquisition (so that not all domain knowledge is provided by experts).

In particular, 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-processor 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 these 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. In fact, we also discuss results on embedding general independence and interventional constraints on pre-trained TPMs. We then conclude the article with observations about the interplay between tractability, learning and knowledge representation in the context of ethical decision-making.

[1] Incidentally, current computational approaches to machine ethics have attempted either bottom-up or top-down pipelines—the latter using learning and the former using formal languages; see discussions and references in Tolmeijer et al. (2020), Charisi et al. (2017), and Hammond and Belle (2021). There is a need to provide frameworks that bridge reasoning and learning in a computationally attractive way.

At the outset, it should be noted that TPMs applied in the above manner are largely propositional, and thus not yet extended to modalities and norms. Nonetheless, these seems like a good starting point for modelling reasoning robots as part of the larger ecosystem where accountability and responsibility is understood more broadly.

Differences to knowledge-enhanced machine learning

It is worth noting that the overall spirit of our argument may seem to be along the lines of knowledge-enhanced ML, there are subtle differences. Indeed, the argument for unifying logic and learning is well-established, by numerous communities, including inductive logic programming (ILP; Muggleton et al., 2012), statistical relational learning (SRL; De Raedt et al., 2016) and neuro-symbolic AI (Hitzler, 2022). However, it should be noted that these areas are primarily motivated by the need to improving classical ML (Murphy, 2012) or compete directly with it by means of domain knowledge and/or logical structure. We argue instead that there is need for the urgent uptake of actions, moral principles, causality, explanations, mental modeling and agency. (Capturing temporal events is not uncommon in the SRL literature (Tran & Davis, 2008), for example, but such notions need to be treated as first-class citizens; not surprisingly, actions have received more attention (Nitti, 2016; Sanner, 2011)).

With regards to the particular case of TPMs, precisely because it has roots in the uncertainty modelling and SRL communities, there is only preliminary work on machine ethics. Moreover, TPM-related research is also preliminary in relating them to expressive KR languages, especially with epistemic, modal and dynamic operators. So, the results here do not take the desired steps yet, but recent work indicate promising directions in which we could proceed. We illustrate the conceptual overlaps in Fig. 1. There are notable exceptions to this categorization, of course, and so this rough depiction only serves to lump together the main themes of different camps:

(1) ILP and SRL's emphasis on integrating reasoning and learning for leveraging deduction machinery and expert knowledge (Muggleton et al., 2012; De Raedt & Kersting, 2011);
(2) Neuro-symbolic AI's unification of deductive machinery and learning (d'Avila Garcez et al., 2002);
(3) TPMs trace their roots to SRL, among other things, to concretely tackle the intractability of logical and probabilistic reasoning via knowledge compilation (Darwiche, 2002a; Darwiche et al., 2016);

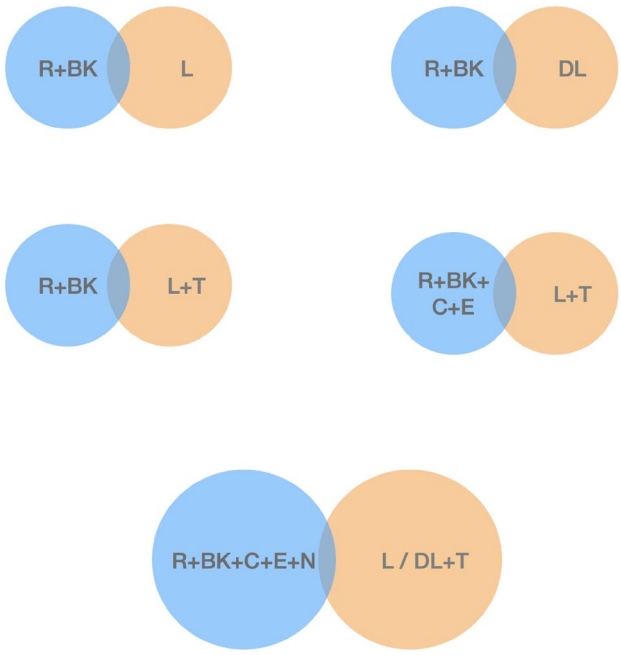

Fig. 1 A rough illustration of existing paradigms. Well-established areas (top row): SRL on the left, and neuro-symbolic AI on the right. Recent development (middle row): TPMs the left, TPMs for ethical and causal AI on the right. Desiderata (bottom row). Abbreviations used: *R* reasoning, *BK* background knowledge, *L* learning/knowledge acquisition, *DL* deep learning, *C* causality, *E* machine ethics, *T* tractability, *N* norms, beliefs and intentions

(4) Recent works using TPMs for ethical and causal AI, discussed below in this article; and
(5) Desiderata explicated above, on unifying above developments with actions, norms, agency, beliefs and other epistemologically grounded concepts.

Long road ahead

It should finally be noted that the research agenda is only taking a small step towards *automation that is accountable*. Computational solutions make strong assumptions about the environment in which the learning and acting happens. Generally, even data collection can amplify positions of privilege, and there are multiple opportunities for failure and misspecification. Orchestrating a framework where this kind of information and knowledge can be communicated between automated systems and humans is not at all obvious, and is an open challenge (Du et al., 2022; Smart et al., 2020). Implementing one or more formal definitions, besides, can either lead to inconsistencies in optimization objectives (Saxena et al., 2019; Smart et al., 2020; Verma & Rubin, 2018; Xiang & Raji, 2019), or might fail altogether on more wholistic fairness ideals such as egalitarianism (Kuppler et al., 2021; Jasso, 1983). A research agenda such as ours does not shed any light on such issues, and is largely

orthogonal. All we suggest is that in so much as complex systems permit the formal specification of environments and epistemological notions, which undoubtedly demands tractable reasoning over possible worlds together with knowledge acquisition, our agenda might offer some promise. Indeed, the two-pronged approach is not advocated as a solution to broader problems, and it is unclear whether abstract models can imbibe cultural and sociopolitical contexts in a straightforward manner. Our agenda 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. This seems like a good starting point for contextual modelling and interacting with reasoning robots. It is then understood that this needs to be a part of the larger ecosystem where accountability and responsibility is understood more broadly (Aplin et al., 2022; Naiseh et al., 2022a, b; Smart et al., 2020).

Key challenges

Given the close connection with many existing areas of AI, such as SRL (as seen above), as well as the difficulty in specifying the delineation from the larger ecosystem in which the formal model is deployed, it is worth articulating what the key challenges are. From existing research on knowledge-enhanced ML (Belle, 2021) and knowledge representation (Lakemeyer et al., 2007), we notice issues such as (taken verbatim from Belle (2021)):

- "What knowledge does a system need to have in advance – i.e., provided by the modeler – versus what can be acquired by observations?" and
- "How does the system generalize from low-level observations to high-level structured knowledge?"

still apply. However, with ethical concerns, a whole range of sociopolitical concerns need to be mapped, or at least suitably interfaced with computational solutions. Chief among them is this:

- Which principles are worth studying computationally? How are they to be formalized, and how can these computational mechanisms interface with notions of accountability and responsibility, broadly construed?

This is arguably a meta-level question, and a deep one. Accuracy measures, communicating explanations, and factoring sources of error as a result of misspecification and distribution drift (Gunning, 2016a; Arrieta et al., 2020; Belle & Papantonis, 2020; Rudin, 2019) need to be ultimately coupled with normative judgements to position the impact of AI systems (Bonnefon et al., 2016; Hurtado et al., 2021; Malle & Scheutz, 2018; Mao & Gratch, 2012; Stilgoe, 2018).

Be that as it may, assuming the scope of automation has been determined, the following questions arise in the very least:

- What sort of environment–actor model is needed for the formalization? From the environment side, do we need concepts such as time, space, observable, controllable and uncontrollable variables? From the actor side, do we need notions for knowledge, belief, beliefs of others, intentions, communication, and social cues (Williams, 2012; Petrick & Foster, 2013)?
- How can the formal principles be embedded in the system? That is, is expert knowledge used for the training of a ML model (e.g., signals for back-propagation in neural networks (Hoernle et al., 2022; Gajowniczek et al., 2020)), or as background knowledge against which all entailments necessarily hold (Muggleton et al., 2012)? If the former, what kind of robustness guarantee is needed to ensure that the signal is not corrupted over the learning epochs? If the latter, how we do prepare against the brittleness of expert knowledge given complex social cues?
- Which principles are expert-level statements (e.g., probabilistic independence between variables) versus those whose weights need to be adjusted as per population data (e.g., learn the probability with which a viral infection spreads in older Asian males)?
- How are different user-level objectives balanced? For example, maximizing accuracy might be at odds with achieving fairness; different notions of fairness may conflict with each other; and explanations might be given to deliberately mislead the end user (Weller, 2019).

These challenges notwithstanding, we think some progress towards the broader program can be made using the following strategy: ethical notions are attempted to be formalized using rich epistemic logics, as seen in recent proposals on blameworthiness, consequentialist and deontological norms (Chockler & Halpern, 2004; Pagnucco et al., 2021), and this is coupled with an account of knowledge acquisition.

Some of the work discussed below are instances of this type of low-hanging fruit, and we hope they provide inspiration for the broader program.

Impact

One recent episode serves to highlight the impact of machine learning models, and necessitates the appropriate application of ethical constraints and de-biasing to ML models. Pro-Publica, a US-based entity specialising in not-for-profit journalism, published an article suggesting that an algorithm widely used to predict the probability of re-offense in criminals was biased against black offenders (Angwin

et al., 2016). The article raised concerns about the fairness and efficacy of the Correctional Offender Management Profiling for Alternative Sanctions (or COMPAS) algorithm, which is widely used in the US justice system. Their article received criticism from both members of the academic community (Flores et al., 2016) and Northpointe, the company who created the algorithm (Dieterich et al., 2016). Their primary complaint concerned the metric used by ProPublica to measure discrimination; the original concerns about racial bias were based mainly on the discrepancy in false positive and false negative rates between black and white offenders. This analysis was critiqued in part because the initial complaint failed to appreciate that the outcome of the algorithm was not a prediction of future behaviour per se, but actually a risk allocation. ProPublica defined a false-positive as any individual considered "high-risk" who did not re-offend, whereas in reality the risk categories were simply an indication of reoffence probability.

This episode illustrates the situation where there is the potential for injustices to arise as a consequence of bias on the basis of training over sensitive factors and variables. Ideally, such factors should be outside the purview of the algorithm's decision making process. Moreover, it is apparent that many criteria used to determine fairness are mutually incompatible (Friedler et al., 2016), and that caution should be used when selecting the criterion for a specific situation. This can lead to significant discrepancies in the interpretation of the same model and its outcomes.

One key dynamic often ignored in mainstream fairness literature is how such decisions might play out in an always-online and continuous operative setting, such in robot-assisted surgery and robotic social care. (See Creager et al. (2020) for a notable exception of a formal nature.) In such applications, designers may not have the opportunity to unplug the system, and reflect on the decision being a proxy for historical behavior versus future actions. If a face recognition system of a robot failed, for example, to detect a human being in the vicinity owing to their skin color, we should expect irrevocable damages, such as crashing into the person, during the robot's operation. Likewise, when the system fails to recognize how its actions could be influenced by the human's sensitive attributes, such as race and gender, we should expect catastrophic scenarios also in applications such as robot-assisted surgery (Hurtado et al., 2021).

When it comes to responsibility and blame, the need for delineating the human–machine boundary is a challenge. In the well-studied infamous trolley problem (Thomson, 1985), a putative agent encounters a runaway trolley headed towards five individuals who are unable to escape from the imminent collision, leading to their deaths. The agent, however, can save them by diverting the trolley to a side track by means of a switch, but at the cost of the death of another individual, who happens to be on this latter track. But the real-world instance of such problems is far more complex. When a self-driving car exhibits problematic behavior, the notion of blame is a multi-faceted issue. In so much as the car leverages machine learning models, service failures can be to a variety of factors from faulty training to an overly optimistic assessment of the error margin. So assigning blame to the 'guilty' party is a tricky affair (Leo & Huh, 2020). Indeed, some have to argued that in light of the fatality from the crash of the Tesla Model S in 2016, responsibility in the presence of ML models is essentially a governance issue (Stilgoe, 2018). Recently funding calls (e.g., `tas.ac.uk`) on trust in autonomous systems also reflect this thinking, and thereby advocating that verifiability and robustness are also facets that feature in this context.

Be that as it may, even if it the case that in practice, the situations encountered by self-driving cars should not involve extreme choices such as whether to save the passenger or the pedestrian, it is still useful for the AI systems to act in line with human values and preferences (Etzioni & Etzioni, 2017). Imbuing such systems with the ability to reason about moral value, blame, and intentionality is one possible step towards this goal.

Two-pronged approach

Our view is that a two-pronged approach is likely needed. On the one hand, we have to draw on philosophical notions and look to formalise them, as attempted by the knowledge representation community. For example, Malle et al. (2014) argue that for blame to emerge, an agent must be perceived as the cause of a negative event. Similarly, Chockler and Halpern (2004) provide an account for the degree of responsibility (versus an 'all or nothing' definition). There are numerous earlier proposals still about social norms, obligation and intentions (Broersen et al., 2001; Georgeff et al., 1998; Jennings, 1993), but they do not necessarily discuss moral factors and blame. To a large extent, nonetheless, these approaches do not focus on the learning of models (actions, beliefs and utilities).

Let us reiterate that, not surprisingly, a large body of work has been considered on ethical artificial intelligence. The topic of fairness has become an increasingly important issue within the field of ML, both in academic circles (Kusner et al., 2017; Zafar et al., 2017; Dwork et al., 2011; Kamishima et al., 2011; Friedler et al., 2016), and more recently in the public and political domains (Angwin et al., 2016; Flores et al., 2016). But as argued previously, much of this literature is focused on the one-shot decision, and very little work has considered the impact of fair behavior in an always-online and continuous operative setting, although there are some exceptions (Hurtado et al., 2021; Creager et al., 2020). As we expect such settings to involve norms

and beliefs, a framework that admits the representation of such epistemic notions alongside fairness considerations seems pertinent.

It is worth remarking that learning from human demonstrations, which is a popular scheme in robotics literature, might be akin to learning patterns from historical data. Thus, just as we expect prejudiced behavior to be embodied and amplified if fairness constraints are not explicitly specified in one-shot decision making, we should also expect that a robot might embody prejudiced behavior from demonstrations. (Admittedly, this is true in principle, as demonstrations are highly controlled settings where explicit prejudices will likely not be on display in an obvious manner; nonetheless, implicit prejudices, such as the absence of training data on underrepresented groups, will likely be present.)

Overall, as far as the fairness literature is concerned, we think incorporating richer models of social interaction are largely lacking. The opposite problem is true for the moral reasoning literature. The latter literature, by design, is built on rich models of agency, beliefs and norms, but is lacking in effective learning mechanisms. For example, the formalization of Halpern and Kleiman-Weiner (2018) is a notable step towards a rigorous proposal for reasoning about causes and blameworthiness. This is essentially based upon prior work done by Chockler and Halpern (2004) and Halpern and Pearl (2005). These frameworks are also related to the intentions model of Kleiman-Weiner et al. which considers predictions about the moral permissibility of actions via influence diagrams (Kim et al., 2018), though unlike our efforts here all of these works are primarily theoretical and there is no emphasis on learning or tractability. Interestingly, the use of tractable architectures for decision-making itself is recent (Bhattacharjya & Shachter, 2012; Melibari et al., 2016). See Hammond and Belle (2021) for detailed discussions. Moreover, undoubtedly, there is a spectrum of ethical issues between fairness and responsibility. Responsible systems need to be fair, but can involve a range of capabilities from social reasoning to verifiable behavior and from error reporting to delegation of decisions (Dignum, 2019).

There is an interesting but somewhat orthogonal development in the related and relevant area of explainable artificial intelligence (Gunning, 2016b). Although the primary emphasis in the area is in exposing a ML model's decision boundary via simplification and rule extraction (Arrieta et al., 2020; Belle & Papantonis, 2020), a number of recent approaches stemming from acting and planning are attempting to build a mental model of the user (Kambhampati, 2020). The idea is that system explanations would be catered to the user's (intuitive) expectation while also gradually refining the system's model of the world. Although tractability and end-to-end learning is not always explicitly addressed, such initiatives fit in squarely with our desiderata.

Progress on tractable learning

As discussed in the previous sections, there are numerous works on capturing complex epistemic and ethical notions, and independently, on learning fair models. However, striking a balance between tractability, learning and reasoning is challenging, and we now discuss a few representative examples where there is emphasis on tractable reasoning and learning. The below works are also very recent, which indicates the preliminary nature of the integration of ethical notions. But as will become clear, it is already bearing fruit, which makes this research direction promising.

In Farnadi et al. (2018), the key observation made is that the standard fairness literature focuses solely on attributes of individuals. A richer language is needed to capture the relationships between individuals and entities, such as social networks and familial connections. Using the statistical relational language of *probabilistic soft logic* (PSL; Bach et al., 2017), they focus on ensuring predictive parity in their models (Dwork et al., 2011; Hardt et al., 2016). PSL is a language for specifying relational syntactic sugar to hinge-loss Markov random fields, which offers tractability by approximation. In particular, certain classes of probabilistic queries in PSL correspond to integer linear programs, which are intractable, but admit convex programming relaxations, which can be solved in polynomial time. It is assumed, however, the domain and the logical rules governing the relationships is specified by an expert, thus the emphasis is on inference as opposed to structure learning.

In Varley and Belle (2021), the construction of a new procedure for pre-processing data is proposed to ensure fair training. That is, first the proposal identifies subsets of mutually independent variables within a training set by leveraging the tractable learning regime of sum–product networks (SPNs; Gens & Domingos, 2013). This allows the technique to identify a collection of 'safe' variables, where the contribution of the protected attribute is removed. This way the pre-processed data can be used to train a fair model using any ML approach.

A natural direction to consider here is whether declarative knowledge and acquired structure can be interleaved, which is an important theme in SRL (De Raedt et al., 2016). Expert knowledge is especially interesting in the fairness literature as it allows us to flexibly define the discrimination patterns of interest. The recent work of Papantonis and Belle (2021) allows SPNs to be trained over prior probabilistic and interventional constraints. Interestingly, in very recent work, Choi et al. (2020) study the implementation of fairness by encoding independence constraints directly when training circuits. Thus, this work indirectly shows how the strands on fairness and constraints could be unified.

In a related context, the tractable model of Kisa et al. (2014), so-called probabilistic sentential decision diagrams (PSDDs), allows for the specification of logical rules, for example, and certain kinds of probabilistic dependencies. Leveraging that, Choi et al. (2020) propose the learning the fair PSDDs by encoding the independence assumptions of a fair distribution as prior knowledge. Such ideas, as mentioned earlier, broadly align with the notion of knowledge-enhanced ML (Cozman & Munhoz, 2021; Belle, 2017, 2020b), where a bridge between symbolic logic and ML is suggested for data efficiency, among other reasons.

The work of Hammond and Belle (2021) deviates from this emphasis on fairness in instead focusing on blameworthiness, as introduced by Halpern and Kleiman-Weiner (2018). The idea is that the causal model (the structural equations) is expected from the expert, by the learning of the probability distributions governing action outcomes, as well as the cost of the actions (which ultimately determines the least blameworthy course of action) is obtained from data. This leads to a hybrid (between data-driven and rule-based methods) computational framework for moral reasoning, which utilizes the specification of causal models, and at the same time exploits many of the desirable properties of PSDDs (such as tractability, semantically meaningful parameters, and the ability to be both learnt from data and include logical constraints). They show that the models in their experiments are reasonable representations of the distributions over the moral scenarios that they are learnt from. Moreover, the learnt utility functions are able to match human preferences with high accuracy using very little data. This leads to blameworthiness reasoning that is, prima facie, in line with human intuitions.

Beyond these, a number of directions are relevant, that hint at both conceptual as well as practical connections with the development above. In Papantonis and Belle (2022), it is shown that TPMs, along with decision trees, Bayesian network classifiers and random forests are essentially multilinear models. This immediately leads to an effective scheme for generating counterfactual explanations (Wachter et al., 2017), including with diversity constraints (Mothilal et al., 2020), the latter having been previously explored only for differentiable models. It turns out that explanations of this type can be given a distinctly logical interpretation: in explainable AI, we are interested in selecting data points with particular properties; for example, with counterfactuals, we are after a point whose label is the opposite of the one considered. By expressing the input–output behavior of classifiers over discrete features as Boolean theories, we can provide a Boolean formula characterizing desired points and that is precisely the explanation (Darwiche & Marquis, 2021).

On the topic of causality, in Zečević et al. (2021), it is shown that we might train TPMs directly on interventional distributions, allowing for effective inference from such distributions. In that regard, Darwiche (2022b) considers a more comprehensive exploration of how TPMs could serve as a scalable and powerful vehicle for causal reasoning.

Incidentally, TPMs are also being explored for a range of computational challenges from other disciplines. For example, Treiber et al. (2020) explore privacy-preserving machine learning using SPNs. In Galindez Olascoaga et al. (2021), analogous to the bespoke computation of deep learning on GPUs, hardware-specific strategies for TPM inference is investigated. In Huang et al. (2021), the classical simulation of quantum algorithms is explored using circuits. These explorations suggest that TPMs might serve as a common computational substrate for several components in an AI system, perhaps leading to deeper interoperability.

While the related literature discussed above pertain mostly to inference with propositional languages, let us briefly comment on developments on the logical expressiveness side. For concreteness, we center this discussion around probabilistic logic programs (and ProbLog De Raedt et al. (2007), in particular).

As discussed before, the compilation of ProbLog to TPMs is well-understood (Fierens et al., 2011a). This has led to various exciting extensions of ProbLog that also rest on circuits for reasoning. For example, DeepProbLog (Manhaeve et al., 2018) integrates low-level concepts obtained from deep learning pipelines with symbolic reasoning, the latter attained through logic programming machinery. In Vennekens et al. (2010), interventions and counterfactuals defined over structural equations are unified with ProbLog. In Smith et al. (2022), recognizing a user or agent's intent is captured in ProbLog. Independently of these developments, there is existing work on using circuits for epistemic logics (Bienvenu et al., 2010), and the use of circuits of problem classes beyond NP (Darwiche et al., 2016, 2018), including modal reasoning. It is therefore not inconceivable that such machinery could be further unified with epistemic extensions of logic programs (Cabalar et al., 2020; Wang & Zhang, 2005), including those supporting nested probabilistic beliefs (Belle & Levesque, 2015).

Discussion and conclusions

There are altogether three takeaways articulated in this article:

(1) In so much as computational machinery can be applied to ethical concerns in complex AI applications, a model for tractable learning together with a reasoning module

for epistemological notions, actions and effects is fundamental.

(2) The only computational model available for the latter is derived from epistemic logic and its variants (including state-based models, belief-level planning frameworks and expressive dynamic modal logics (Kaelbling & Lozano-Pérez, 2013; Kaelbling et al., 1998; Sanner & Kersting, 2010; Belle & Lakemeyer, 2017)).

(3) An emerging paradigm for tractable probabilistic learning is based on knowledge compilation target languages such as propositional circuits.

In sum, TPMs offers the most compelling computational framework for tractability + logical reasoning about possible worlds + knowledge acquisition, and hence it is a worthwhile starting point for the endeavor.

As mentioned before, the tractable learning paradigm is in its early years. And, at least as far as capturing a broad range of knowledge representation languages is concerned, there is altogether less emphasis on mental modeling and agency. (First-order expressiveness is yet another dimension for allowing richness in specifications, as already admitted by some relational probabilistic models (Getoor & Taskar, 2007), lifted approaches (Van den Broeck, 2011) and proposals with an explicit causal theory such as Salimi et al. (2020) and Vennekens et al. (2010)). In contrast, readers may want to consult discussions in Kambhampati (2020) and Hammond and Belle (2021) on knowledge representation approaches where a more comprehensive model of the environment and its actors is considered, but where knowledge acquisition and learning are either ignored or dealt with in careful, limited ways.

It is also interesting to note that although many expressive languages (Van den Broeck et al., 2010; Fierens et al., 2011b) 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. Indeed, the literature discussed above do take steps of this sort. For example, Hammond and Belle (2021) expect an expert to provide the causal model, but learns the probabilities and utilities from data. Analogously, Choi et al. (2020) and Papantonis and Belle (2021) expect the provision of independence independence assumptions of a distribution, but the underlying probabilistic model and distribution is learnt from data. There is also an emerging literature on abstraction, and how high-level concepts might be mapped to low-level data (Beckers & Halpern, 2019; Belle, 2020a): adapting that literature in conjunction with table learning might be promising too.

It is worth noting that there are other paradigms of tractable learning, including but not limited to those based on the probably approximately correct (PAC) learning semantics (Juba, 2013). These have recently enjoyed extensions to expressive logical languages, including non-trivial fragments of first-order logic (Belle & Juba, 2019; Mocanu et al., 2020). These might serve as an alternative paradigm to TPMs in service of our overall objectives, which could be an interesting direction for the future.

Let us conclude with key observations about the results discussed. The 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 (Batusov & Soutchanski, 2018) and multiple agents (Ghaderi et al., 2007). 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 TPMs (or other structures with analogous properties) with such rich formal systems? As mentioned, it is known that certain probabilistic logical theories (Fierens et al., 2011b) can be reduced to such structures, so perhaps gentle extensions to those theories (and as well as reversing the pipeline) might suggest ways to integrate causal epistemic models and tractable learning.

We have repeatedly emphasized the notion of tractability as a desirable characteristic for the computational model to have. But what if certain ethical notions are provably intractable? Should we then only settle for approximate notions that are provably tractable, or abandon the issue of tractability completely? There is no clear answer to this, and it might depend on the application at hand. Perhaps the situation is not dissimilar to the encountering of hard computational problems in the real world. Many computational tasks can be encoded as satisfiability and validity problems in propositional and higher-order logics, but satisfiability is NP-complete already in finitary propositional logic. Although a great deal of attention has been dedicated to identifying when and where problems requiring exponential time emerge (Mitchell et al., 1992), very many real-world problems get solved, and in real-time no less (Barrett et al., 2009; Kautz & Selman, 1992), prompting the search for technology that goes after

much harder classes beyond NP-complete (Gomes et al., 2009; Ermon et al., 2013). So too might be our encounter with automating machine ethics: either certain intractable problems might be solved approximately, or fragments and restrictions may be applied so as to be solved exactly, or the seriousness of the situation may demand an exact solution regardless of the computational resources needed. What is clear, however, is that agents need to respond to signals and observations from the real-world, so tractability and scalability in knowledge acquisition and reasoning are important considerations.

Beyond that technical front, as discussed in our introduction, much work remains to be done, of course, in terms of delineating automated decision-making from delegation and notions of accountability (Dignum, 2019; Crootof et al., 2022). 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 (Crawford, 2021a, b). Orchestrating a framework where this kind of information and knowledge can be communicated back and forth between automated systems and stakeholders 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. We hope that this type of expressiveness offers additional protection, control, explanation and normative reasoning during the deployment of complex systems with ML components.

Acknowledgements An earlier version was delivered as a keynote at *27th International Conference on Conceptual Structures*, 2022, Münster, Germany, and also presented at the *Critical Perspectives on Artificial Intelligence Ethics Conference*, 2020, Edinburgh, UK.

Funding This research was partly supported by a Royal Society University Research Fellowship, UK, and partly supported by a Grant from the UKRI Strategic Priorities Fund, UK to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1, 2020–2024).

References

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence, 258*, 66–95.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. Benton Institute.

Aplin, T., Schafer, B., & Li, P. (2022). *Trustworthy autonomous systems hub and TAS node on regulation and governance (2021) artificial intelligence and IP: Copyright and patents: A call for evidence from Intellectual Property Office*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Bach, S. H., Broecheler, M., Huang, B., & Getoor, L. (2017). Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research, 18*, 1–67.

Barrett, C., Sebastiani, R., Seshia, S. A., & Tinelli, C. (2009). Chapter 26: Satisfiability modulo theories. In *Handbook of satisfiability* (pp. 825–885). IOS Press.

Batusov, V., & Soutchanski, M. (2018). Situation calculus semantics for actual causality. In *Proceedings of the AAAI conference on artificial intelligence*, 2018 (Vol. 32).

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the AAAI conference on artificial intelligence*, 2019 (Vol. 33, pp. 2678–2685).

Belle, V. (2017). Logic meets probability: Towards explainable AI systems for uncertain worlds. In *IJCAI*, 2017 (pp. 5116–5120).

Belle, V. (2020a). Abstracting probabilistic models: Relations, constraints and beyond. *Knowledge-Based Systems, 199*, 105976.

Belle, V. (2020b). Symbolic logic meets machine learning: A brief survey in infinite domains. In *International conference on scalable uncertainty management*, 2020 (pp. 3–16). Springer.

Belle, V. (2021). Logic meets learning: From Aristotle to neural networks. In *Neuro-symbolic artificial intelligence: The state of the art* (pp. 78–102). IOS Press.

Belle, V., & Juba, B. (2019). Implicitly learning to reason in first-order logic. *Advances in neural information processing systems*, 2019 (Vol. 32).

Belle, V., & Lakemeyer, G. (2017). Reasoning about probabilities in unbounded first-order dynamical domains. In *IJCAI*, 2017.

Belle, V., & Levesque, H. J. (2015). ALLEGRO: Belief-based programming in stochastic dynamical domains. In *IJCAI*, 2015.

Belle, V., & Papantonis, I. (2020). Principles and practice of explainable machine learning. arXiv preprint. arXiv:2009.11698

Bhattacharjya, D., & Shachter, R. D. (2012). Evaluating influence diagrams with decision circuits. arXiv preprint. arXiv:1206.5257

Bienvenu, M., Fargier, H., & Marquis, P. (2010). Knowledge compilation in the modal logic S5. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Brachman, R. J., Levesque, H. J., & Reiter, R. (1992). *Knowledge representation*. MIT Press.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., & van der Torre, L. (2001). The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents*, 2001 (pp. 9–16).

Cabalar, P., Fandinno, J., Garea, J., Romero, J., & Schaub, T. (2020). eclingo: A solver for epistemic logic programs. *Theory and Practice of Logic Programming, 20*(6), 834–847.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A. F., & Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint. arXiv:1703.04741

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research, 22*, 93–115.

Choi, Y., Dang, M., & Van den Broeck, G. (2020). Group fairness by probabilistic modeling with latent fair decisions. arXiv preprint. arXiv:2009.09031

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Cozman, F. G., & Munhoz, H. N. (2021). Some thoughts on knowledge-enhanced machine learning. *International Journal of Approximate Reasoning, 136,* 308–324.

Crawford, K. (2021a). *The atlas of AI*. Yale University Press.

Crawford, K. (2021b). The hidden costs of AI. *New Scientist, 249*(3327), 46–49.

Creager, E., Madras, D., Pitassi, T., & Zemel, R. (2020). Causal modeling for fairness in dynamical systems. In *International conference on machine learning*, 2020 (pp. 2185–2195). PMLR.

Crootof, R., Kaminski, M. E., & Price, W. N., II. (2022). Humans in the loop. *Vanderbilt Law Review*. https://doi.org/10.2139/ssrn.4066781.

Darwiche, A. (2002a). A logical approach to factoring belief networks. In *Proceedings of the 8th international conference on principles of knowledge representation and reasoning*, 2002 (pp. 409–420).

Darwiche, A. (2022b). Causal inference using tractable circuits. arXiv preprint. arXiv:2202.02891

Darwiche, A., Marques-Silva, J., & Marquis, P. (2016). Preface: The beyond NP workshop. In *Beyond NP, papers from the 2016 AAAI workshop*, Phoenix, Arizona, USA, February 12, 2016.

Darwiche, A., & Marquis, P. (2021). On quantifying literals in Boolean logic and its applications to explainable AI. *Journal of Artificial Intelligence Research, 72*, 285–328.

Darwiche, A., Marquis, P., Suciu, D., & Szeider, S. (2018). Recent trends in knowledge compilation (Dagstuhl seminar 17381). In *Dagstuhl reports* (Vol. 7). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik

d'Avila Garcez, A. S., Broda, K., Gabbay, D. M., et al. (2002). *Neural-symbolic learning systems: Foundations and applications*. Springer.

De Raedt, L., & Kersting, K. (2011). Statistical relational learning. In *Encyclopedia of machine learning* (pp. 916–924). Springer.

De Raedt, L., Kersting, K., Natarajan, S., & Poole, D. (2016). Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 10*(2), 1–189.

De Raedt, L., Kimmig, A., & Toivonen, H. (2007). ProbLog: A probabilistic prolog and its application in link discovery. In *Proceedings of IJCAI*, 2007 (pp. 2462–2467).

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical Report. Northpointe.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., et al. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint. arXiv:1711.01134

Du, X., Legastelois, B., Ganesh, B., Rajan, A., Chockler, H., Belle, V. Anderson, S., & Ramamoorthy, S. (2022). Vision checklist: Towards testable error analysis of image models to help system designers interrogate model capabilities. arXiv preprint. arXiv:2201.11674

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. In *Innovations in theoretical computer science conference*, 2011.

Ermon, S., Gomes, C. P., Sabharwal, A., & Selman, B. (2013). Embed and project: Discrete sampling with universal hashing. In *NIPS*, 2013 (pp. 2085–2093).

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418.

Fagin, R., Moses, Y., Halpern, J. Y., & Vardi, M. Y. (2003). *Reasoning about knowledge*. MIT Press.

Farnadi, G., Babaki, B., & Getoor, L. (2018). Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*, 2018 (pp. 108–114).

Fierens, D., Van den Broeck, G., Thon, I., Gutmann, B., & De Raedt, L. (2011a). Inference in probabilistic logic programs using weighted CNF's. In *UAI*, 2011 (pp. 211–220).

Fierens, D., Van den Broeck, G., Thon, I., Gutmann, B., & De Raedt, L. (2011b). Inference in probabilistic logic programs using weighted CNF's. In *Proceedings of UAI*, 2011 (pp. 211–220).

Flores, A. W., Lowenkamp, C., & Bechtel, K. (2016). False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. And it's biased against blacks". *Federal Probation, 80*(2).

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (IM) possibility of fairness. arXiv preprint. arXiv:1609.07236.

Gajowniczek, K., Liang, Y., Friedman, T., Zabkowski, T., & Van den Broeck, G. (2020). Semantic and generalized entropy loss functions for semi-supervised deep learning. *Entropy, 22*(3), 334.

Galindez Olascoaga, L. I., Meert, W., & Verhelst, M. (2021). Hardware-aware probabilistic circuits. In *Hardware-aware probabilistic machine learning models* (pp. 81–110). Springer.

Gens, R., & Domingos, P. (2013). Learning the structure of sum–product networks. In *International conference on machine learning*, 2013.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. (1998). The belief–desire–intention model of agency. In *International workshop on agent theories, architectures, and languages*, 1998 (pp. 1–10). Springer.

Getoor, L., & Taskar, B. (2007). *Introduction to statistical relational learning (adaptive computation and machine learning)*. MIT Press.

Ghaderi, H., Levesque, H., & Lespérance, Y. (2007). Towards a logical theory of coordination and joint ability. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 2007 (pp. 1–3).

Gomes, C. P., Sabharwal, A., & Selman, B. (2009). Model counting. In *Handbook of satisfiability*. IOS Press.

Gunning, D. (2016a). Explainable artificial intelligence (XAI). Technical Report, DARPA/I20.

Gunning, D. (2016b). *Explainable artificial intelligence (XAI)—DARPA-BAA-16-53*. Defense Advanced Research Projects Agency.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Halpern, J. Y. (2017). *Reasoning about uncertainty*. MIT Press.

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility.

In *Proceedings of the 32nd AAAI conference on artificial intelligence*, 2018 (pp. 1853–1860).
Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887.
Hammond, L., & Belle, V. (2021). Learning tractable probabilistic models for moral responsibility and blame. *Data Mining and Knowledge Discovery, 35*(2), 621–659.
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *International conference on neural information processing systems*, 2016.
Hitzler, P. (2022). *Neuro-symbolic artificial intelligence: The state of the art*. IOS Press.
Hoernle, N., Karampatsis, R. M., Belle, V., & Gal, K. (2022). MultiplexNet: Towards fully satisfied logical constraints in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2022 (Vol. 36, pp. 5700–5709).
Huang, Y., Holtzen, S., Millstein, T., Van den Broeck, G., & Martonosi, M. (2021). Logical abstractions for noisy variational quantum algorithm simulation. In *Proceedings of the 26th ACM international conference on architectural support for programming languages and operating systems*, 2021 (pp. 456–472).
Hurtado, J. V., Londoño, L., & Valada, A. (2021). From learning to relearning: A framework for diminishing bias in social robot navigation. arXiv preprint. arXiv:2101.02647
Jasso, G. (1983). Fairness of individual rewards and fairness of the reward distribution: Specifying the inconsistency between the micro and macro principles of justice. *Social Psychology Quarterly, 46*(3), 185–199.
Jennings, N. R. (1993). Specification and implementation of a belief–desire–joint–intention architecture for collaborative problem solving. *International Journal of Intelligent and Cooperative Information Systems, 2*(03), 289–318.
Juba, B. (2013). Implicit learning of common sense for reasoning. In *Twenty-third international joint conference on artificial intelligence*, 2013.
Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence, 101*(1–2), 99–134.
Kaelbling, L. P., & Lozano-Pérez, T. (2013). Integrated task and motion planning in belief space. I. *Journal of Robotic Research, 32*(9–10), 1194–1227.
Kambhampati, S. (2020). Challenges of human-aware AI systems. *AI Magazine, 41*(3), 3–17.
Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *Conference on data mining*, 2011.
Kautz, H., & Selman, B. (1992). Planning as satisfiability. In *ECAI '92: Proceedings of the 10th European conference on Artificial intelligence*, 1992 (pp. 359–363). Wiley.
Khandani, A., Kim, J., & Lo, A. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance, 34*(11), 2767–2787.
Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J. B., & Rahwan, I. (2018). A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*, 2018 (pp. 197–203).
Kisa, D., Van den Broeck, G., Choi, A., & Darwiche, A. (2014). Probabilistic sentential decision diagrams. In *Proceedings of the 14th international conference on principles of knowledge representation and reasoning*, 2014 (pp. 558–567).
Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2021). Distributive justice and fairness metrics in automated decision-making: How much overlap is there? arXiv preprint. arXiv:2105.01441
Kusner, M., Loftus, J., Russel, C., & Silva, R. (2017). Counterfactual fairness. In *Neural information processing systems*, 2017.
Lakemeyer, G., & Levesque, H. J. (2007). Cognitive robotics. In *Handbook of knowledge representation* (pp. 869–886). Elsevier.
Leo, X., & Huh, Y. E. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior, 113*, 106520.
Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.
Malle, B. F., & Scheutz, M. (2018). Learning how to behave: Moral competence for social robots. In: *Handbuch Maschinenethik* (pp. 1–24).
Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). DeepProbLog: Neural probabilistic logic programming. *Advances in neural information processing systems*, 2018 (Vol. 31).
Mao, W., & Gratch, J. (2012). Modeling social causality and responsibility judgement in multi-agent interactions. *Journal of Artificial Intelligence Research, 44*, 223–273.
Melibari, M., Poupart, P., & Doshi, P. (2016). Sum–product-max networks for tractable decision making. In *IJCAI*, 2016.
Mitchell, D. G., Selman, B., & Levesque, H. J. (1992). Hard and easy distributions of SAT problems. In *Proceedings of AAAI*, 1992 (pp. 459–465).
Mocanu, I. G., Belle, V., & Juba, B. (2020). Polynomial-time implicit learnability in SMT. In *ECAI 2020*, 2020 (pp. 1152–1158). IOS Press.
Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020 (pp. 607–617).
Muggleton, S., De Raedt, L., Poole, D., Bratko, I., Flach, P., Inoue, K., & Srinivasan, A. (2012). ILP turns 20. *Machine Learning, 86*(1), 3–23.
Murphy, K. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.
Naiseh, M., Bentley, C., & Ramchurn, S. (2022). Trustworthy autonomous systems (TAS): Engaging TAS experts in curriculum design. In *2022 IEEE global engineering education conference (EDUCON)*, 2022 (pp. 901–905). IEEE.
Naiseh, M., Bentley, C., Ramchurn, S., Williams, E., Awad, E., & Alix, C. (2022). Methods, tools and techniques for trustworthy autonomous systems (TAS) design and development. In *Companion of the 2022 ACM SIGCHI symposium on engineering interactive computing systems*, 2022 (pp. 66–69).
Nitti, D. (2016). Hybrid probabilistic logic programming. PhD Thesis, KU Leuven.
Pagnucco, M., Rajaratnam, D., Limarga, R., Nayak, A., & Song, Y. (2021). Epistemic reasoning for machine ethics with situation calculus. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society*, 2021 (pp. 814–821).
Papantonis, I., & Belle, V. (2021). Closed-form results for prior constraints in sum–product networks. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2021.644062.
Papantonis, I., & Belle, V. (2022). Principled diverse counterfactuals in multilinear models. arXiv preprint. arXiv:2201.06467
Petrick, R. P. A., & Foster, M. (2013). Planning for social interaction in a robot bartender domain. In *Proceedings of ICAPS*, 2013, Rome, Italy (pp. 389–397).
Reiter, R. (2001). *Knowledge in action: Logical foundations for specifying and implementing dynamical systems*. MIT Press.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Salimi, B., Parikh, H., Kayali, M., Getoor, L., Roy, S.. & Suciu, D. (2020). Causal relational learning. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, 2020 (pp. 241–256).

Sanner, S. (2011). Relational dynamic influence diagram language (RDDL): Language description. Technical Report. Australian National University.

Sanner, S., & Kersting, K. (2010). Symbolic dynamic programming for first-order POMDPs. In *Proceedings of AAAI*, 2010 (pp. 1140–1146).

Sardina, S., De Giacomo, G., Lespérance, Y., & Levesque, H. J. (2006). On the limits of planning over belief states under strict uncertainty. In *KR*, 2006 (Vol. 6, pp. 463–471).

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, 2019 (pp. 99–106).

Smart, A., James, L., Hutchinson, B., Wu, S., & Vallor, S. (2020). Why reliabilism is not enough. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, 2020.

Smith, G. B., Belle, V., & Petrick, R. (2022). Intention recognition with ProbLog. *Frontiers in Artificial Intelligence, 5,* 75.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science, 48*(1), 25–56.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys, 53*(6), 1–38.

Tran, S. D., & Davis, L. S. (2008). Event modeling and recognition using Markov logic networks. In *Proceedings of ECCV*, 2008 (pp. 610–623).

Treiber, A., Molina, A., Weinert, C., Schneider, T., & Kersting, K. (2020). CryptoSPN: Privacy-preserving sum–product network inference. arXiv preprint. arXiv:2002.00801

Van den Broeck, G. (2011). On the completeness of first-order knowledge compilation for lifted probabilistic inference. In *NIPS*, 2011 (pp. 1386–1394).

Van den Broeck, G., Thon, I., van Otterlo, M., & De Raedt, L (2010). DTProbLog: A decision-theoretic probabilistic prolog. In *Proceedings of the twenty-fourth AAAI conference on artificial intelligence*, AAAI'10, 2010 (pp. 1217–1222). AAAI Press.

Varley, M., & Belle, V. (2021). Fairness in machine learning with tractable models. *Knowledge-Based Systems, 215*, 106715.

Vennekens, J., Bruynooghe, M., & Denecker, M. (2010). Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In *European workshop on logics in artificial intelligence*, 2010 (pp. 313–325). Springer.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM international workshop on software fairness (FairWare)*, 2018 (pp. 1–7). IEEE.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*, 841.

Wang, K., & Zhang, Y. (2005). Nested epistemic logic programs. In *International conference on logic programming and nonmonotonic reasoning*, 2005 (pp. 279–290). Springer.

Weller, A. (2019). Transparency: motivations and challenges. In *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 23–40). Springer.

Williams, M.-A. (2012). Robot social intelligence. In *ICSR*, 2012 (pp. 45–55).

Xiang, A., & Raji, I. D. (2019). On the legal compatibility of fairness definitions. arXiv preprint. arXiv:1912.00761

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *International conference on World Wide Web*, 2017.

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research, 64*, 243–252.

Zečević, M., Dhami, D., Karanam, A., Natarajan, S., & Kersting, K. (2021). Interventional sum–product networks: Causal inference with tractable probabilistic models. *Advances in neural information processing systems*, 2021 (Vol. 34).