TPLP **23** (4): 865–883, 2023.

doi:10.1017/S1471068423000157 First published online 19 July 2023

Toward A Logical Theory Of Fairness and Bias

VAISHAK BELLE
University of Edinburgh & Alan Turing Institute, UK
(*e-mail:* vbelle@ed.ac.uk)

submitted 17 May 2023; revised 10 June 2023; accepted 12 June 2023

Abstract

Fairness in machine learning is of considerable interest in recent years owing to the propensity of algorithms trained on historical data to amplify and perpetuate historical biases. In this paper, we argue for a formal reconstruction of fairness definitions, not so much to replace existing definitions but to ground their application in an epistemic setting and allow for rich environmental modeling. Consequently we look into three notions: fairness through unawareness, demographic parity and counterfactual fairness, and formalize these in the epistemic situation calculus.

KEYWORDS: logic, fairness, bias, situation calculus, knowledge, action

1 Introduction

Machine Learning techniques have become pervasive across a range of different applications, and are the source of considerable excitement but also debate. For example, they are now widely used in areas as disparate as recidivism prediction, consumer credit-risk analysis and insurance pricing (Chouldechova 2017; Khandani *et al.* 2010). In some of these applications, the prevalence of machine learning techniques has raised concerns about the potential for learned algorithms to become biased against certain groups. This issue is of particular concern in cases when algorithms are used to make decisions that could have far-reaching consequences for individuals (for example in recidivism prediction) (Chouldechova 2017; Angwin *et al.* 2016). Attributes which the algorithm should be "fair" with respect to are typically referred to as *protected* attributes. The values to these are often hidden from the view of the decision maker (whether automated or human). There are multiple different potential fields that might qualify as protected attributes in a given situation, including ethnicity, sex, age, nationality and marital status (Zemel *et al.* 2013). Ideally, such attributes should not affect any prediction made by "fair" algorithms. However, even in cases where it is clear which attributes should be protected, there are multiple (and often mutually exclusive) definitions of what it means for an algorithm to be unbiased with respect to these attributes, and there is disagreement within the academic community on what is most appropriate (Dwork *et al.* 2011; Kusner *et al.* 2017; Zafar *et al.* 2017a).

However, even amid pressing concerns that algorithms currently in use may exhibit racial biases, there remains a lack of agreement about how to effectively implement fairness, given the complex socio-technical situations that such applications are deployed in and the background knowledge and context needed to assess the impact of outcomes (e.g. denying a loan to someone in need).

To address such issues broadly, an interesting argument has been championed by the symbolic community: by assuming a rich enough understanding of the application domain, we can encode machine ethics in a formal language. Of course, with recent advances in statistical relational learning, neuro-symbolic AI and inductive logic programming (Raedt *et al.* 2016; Muggleton *et al.* 2012), it is possible to integrate low-level pattern recognition based on sensory data with high-level formal specifications. For example, the *Hera* project (Lindner *et al.* 2017) allows for the implementation of several kinds of (rule-based) moral theory to be captured. *Geneth* (Anderson and Anderson 2014) uses inductive logic generalized moral principles from the judgments of ethicists about particular ethical dilemmas, with the system's performance being evaluated using an *ethical Turing test*. On the formalization side, study of moral concepts has long been a favored topic in the knowledge representation community (Conway and Gawronski 2013; Alexander and Moore 2016; Czelakowski 1997; Hooker and Kim 2018), that can be further coupled against notions of beliefs, desires and intentions (Broersen *et al.* 2001; Georgeff *et al.* 1998). Finally, closer to the thrust of this paper, (Pagnucco *et al.* 2021) formalize consequentialist and deontological ethical principles in terms of "desirable" states in the epistemic situation calculus, and (Classen and Delgrande 2020) formalize obligations using situation calculus programs.

2 Contributions

Our thesis, in essence, is this: complementing the vibrant work in the ML community, it is worthwhile to study ethical notions in formal languages. This serves three broad objectives:

A. We can identify what the system needs to know versus what is simply true (Reiter 2001b; Halpern and Moses 2014) and better articulate how this knowledge should impact the agent's choices. It is worth remarking that epistemic logic has served as the foundation for investigating the impact of knowledge on plans and protocols (Levesque 1996; Lespérance *et al.* 2000; Halpern *et al.* 2009).

B. We implicitly understand that we can further condition actions against background knowledge (such as ontologies and databases), as well as notions such as intentions and obligations (Sardina and Lespérance 2010).

C. We can position the system's actions not simply as a single-shot decision or prediction, as is usual in the ML literature, but as a sequence of complex events that depend on observations and can involve loops and recursion: that is, in the form of programs (Levesque *et al.* 1997).

It would beyond the scope of a single paper to illustrate the interplay between the three objectives except in some particular application scenario. Thus, we focus on the interplay between A and C in the sense of advocating a "research agenda," rather than a single technical result, or a demonstration of a single application. In particular, what

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

we seek to do is a formal reconstruction of some fairness definitions, not so much to replace existing definitions but to ground their application in an epistemic, dynamic setting. Consequently we look into three notions: fairness through unawareness, demographic parity and counterfactual fairness, and formalize these in the epistemic situation calculus (Scherl and Levesque 2003; Lakemeyer and Levesque 2011). In particular, our contributions are as follows:

- Consider the notion of fairness through unawareness (FTU) in machine learning. Here, a "fair" classifier is one that predicts outputs by not using any information about protected attributes. In a dynamic setting, imagine a (virtual or physical) robot that is acting in service of some objective ϕ. For example, in a loan setting, which is classically treated as a static model in machine learning, we can expect intelligent automated agents to carry out many operations: check the yearly budget of the bank to determine the total amount to be loaned, rank applicants based on risk, determine the impact of recession, and ultimately synthesize a plan to achieve ϕ (loan approval), but by virtue of FTU, it should never be the case that the agent has had access to protected information. In this paper, we provide a simple but general definition to capture that idea, in a manner that distinguishes what is true from what is known by the agent.
- Analogously, consider the notion of demographic parity (DP). It is understood as a classifier that is equally likely to make a positive prediction regardless of the value of the protected attribute. For example, the proportion of men who are granted loans equals the proportion of women granted loans. So, if $\phi(x)$ is the granting of a loan to individual x, how do we capture the notion that the agent has synthesized a plan that achieves $\phi(x)$ for both males as well as females? What would it look like for planning agents that want to conform to both FTU and DP? What if, instead of DP, we wished to only look at those granted loans, and among this group, we did not want the classifier to discriminate based on the individual's gender? For all these cases, we provide definitions in terms of the agent's mental state and action sequences that the agent knows will achieve $\phi(x)$ (Levesque 1996).
- Finally, counterfactual fairness insists that the prediction should not differ if the individual's protected attributes take on a different value. For a planning agent to ensure this, we would need to make sure that *deleting* facts about the current value for an individual x's protected attribute and *adding* a different value still achieves $\phi(x)$ after the sequence. We characterize this using the notion of *forgetting* because we permit, in general, any arbitrary first-order theory for the initial knowledge base, and not just a database interpreted under the closed-world assumption.

These definitions can be seen to realize a specification for "fair" cognitive robots: that is, reasoning and planning agents (Lakemeyer and Levesque 2007) that ensure through the course of their acting that, say, they never gain knowledge about the protected attributes of individuals, and guarantee that individuals are not discriminated based on values to these attributes.

It should be clear that our definitions are loosely inspired by the ML notions. And so our formalization do not argue for one definition over another, nor challenge any existing definition. We do, however, believe that studying the effects of these definitions in a dynamic setting provides a richer context to evaluate their appropriateness. Moreover, a

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

formalization such as ours lends itself to various types of implementations. For example, the synthesis of (epistemic) programs and plans (Wang and Zhang 2005; Baral *et al.* 2017; Muise *et al.* 2015; Classen *et al.* 2008; McIlraith and Son 2002) that achieve goals in socio-technical applications in a fair manner is an worthwhile research agenda. Likewise, enforcing fairness constraints while factoring for the relationships between individuals in social networks (Farnadi *et al.* 2018), or otherwise contextualizing attributes against other concepts in a relational knowledge base (Aziz *et al.* 2018; Fu *et al.* 2020) are also worthwhile. By stipulating an account in quantified logic, it becomes possible to further unify such proposals in a dynamic setting.

Logic and fairness. Let us briefly remark on closely related efforts. At the outset, note that although there has been considerable work on formalizing moral rules, there is no work (as far as we are aware) on the formalization of fairness and bias in a *dynamic epistemic* setting, where we need to explicate the interaction between actions, plans and meta-beliefs. However, there is some work that tackles epistemic and logical aspects.

For example, the work of Kawamoto (2019) considers a statistical epistemic logic and its use for the formalization of statistical accuracy as well as fairness, including the criterion of equality of opportunity. There are a few key differences to our work: that work is motivated by a probabilistic reconstruction of prediction systems by appealing to distance measures, and so knowledge is defined in terms of accessibility between worlds that are close enough. The language, moreover, allows for "measurement" variables that are interpreted statistically. In contrast, our account is not (yet) probabilistic, and if our account were to be extended in that fashion, the most obvious version would reason about degrees of belief (Bacchus *et al.* 1999; Belle and Lakemeyer 2017); see Bacchus *et al.* (1996) for discussions on the differences between statistical belief and degrees of belief. Moreover, our account is dynamic, allowing for explicit modalities operators for actions and programs. Consequently, our definitions are about studying how, say, the agent remains ignorant about protected attributes when executing a plan.

Be that as it may, the work of Kawamoto (2019) leads to an account where fairness can be expressed as a logical property using predicates for protected attributes, remarkably similar in spirit to our approach if one were to ignore actions. This should, in the very least, suggest that such attempts are very promising, and for the future, it would be worthwhile to conduct a deeper investigation on how these formalization attempts can be synthesized to obtain a general probabilistic logical account that combines the strength of dynamic epistemic languages and statistical measures. (In a related vein to Kawamoto (2019), Liu and Lorini (2022) seek to axiomatize ML systems for the purpose of explanations in a modal logic.) An entirely complementary effort is the use of logic for verifying fair models (Ignatiev *et al.* 2020), where existing definitions and classifiers are encoded using logical functions and satisfiability modulo theories.

To summarize, all these differ from our work in that we are attempting to understand the interplay between bias, action and knowledge, and not really interested in capturing classifiers as objects in our language. Thus, our work, as discussed above, can be seen as setting the stage for *"fair" cognitive robots*. There is benefit to unifying these streams, which we leave to the future.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

3 A logic for knowledge and action

We now introduce the logic $\mathcal{ES}$ (Lakemeyer and Levesque 2004).[1] The non-modal fragment of $\mathcal{ES}$ consists of standard first-order logic with $=$. That is, connectives $\{\wedge, \forall, \neg\}$, syntactic abbreviations $\{\exists, \equiv, \supset\}$ defined from those connectives, and a supply of variables $\{x, y, \ldots, u, v, \ldots\}$. Different to the standard syntax, however, is the inclusion of (countably many) *standard names* (or simply, names) for both objects and actions $\mathcal{R}$, which will allow a simple, substitutional interpretation for $\forall$ and $\exists$. These can be thought of as special extra constants that satisfy the unique name assumption and an infinitary version of domain closure.

Like in the situation calculus, to model immutable properties, we assume rigid predicates and functions, such as *IsPlant(x)* and *father(x)* respectively. To model changing properties, $\mathcal{ES}$ includes fluent predicates and functions of every arity, such as *Broken(x)* and *height(x)*. Note that there is no longer a situation term as an argument in these symbols to distinguish the fluents from the rigids. For example, $\mathcal{ES}$ also includes distinguished fluent predicates *Poss* and *SF* to model the executability of actions and capture sensing outcomes respectively, but they are unary predicates (i.e. in contrast to the classical situation calculus (Reiter 2001a) because they no longer include situation terms.) Terms and formulas are constructed as usual. The set of ground atoms $\mathcal{P}$ are obtained, as usual, from names and predicates.

There are four modal operators in $\mathcal{ES}$: $[a], \square, \boldsymbol{K}$ and $\boldsymbol{O}$. For any formula α, we read $[a]\alpha, \square\alpha$ and $\boldsymbol{K}\alpha$ as "α holds after a," "α holds after any sequence of actions" and "α is known," respectively. Moreover, $\boldsymbol{O}\alpha$ is to be read as "α is only-known." Given a sequence $\delta = a_1 \cdots a_k$, we write $[\delta]\alpha$ to mean $[a_1] \cdots [a_k]\alpha$.

In classical situation calculus parlance, we would use $[a]\alpha$ to capture successor situations as properties that are true after an action in terms of the current state of affairs. Together with the $\square$ modality, which allows to capture quantification over situations and histories, basic action theories can be defined. Like in the classical approach, one is interested in the entailments of the basic action theory.

Semantics. Recall that in the simplest setup of the possible-worlds semantics, worlds mapped propositions to $\{0, 1\}$, capturing the (current) state of affairs. $\mathcal{ES}$ is based on the very same idea, but extended to dynamical systems. So, suppose a world maps $\mathcal{P}$ and $\mathcal{Z}$ to $\{0, 1\}$.[2] Here, $\mathcal{Z}$ is the set of all finite sequences of action names, including the empty sequence $\langle\rangle$. Let $\mathcal{W}$ be the set of all worlds, and $e \subseteq \mathcal{W}$ be the *epistemic state*. By a *model*, we mean a triple (e, w, z) where $z \in \mathcal{Z}$. Intuitively, each world can be thought of as a situation calculus tree, denoting the properties true initially but also after every sequence of actions. $\mathcal{W}$ is then the set of all such trees. Given a triple (e, w, z), w denotes the real world, and z the actions executed so far.

[1] Our choice of language may seem unusual, but it is worth noting that this language is a modal syntactic variant of the classical epistemic situation that is better geared for reasoning about knowledge (Lakemeyer and Levesque 2011). But more importantly, it can be shown that reasoning about actions and knowledge reduces to first-order reasoning via the so-called regression and representation theorems (Lakemeyer and Levesque 2004). (For space reasons, we do not discuss such matters further here.) There are, of course, many works explicating the links between the situation calculus and logic programming; see, for example, Lee and Palla (2012). See also works that link the situation calculus to planning, such as Classen *et al.* (2008); Belle (2022); Sardina *et al.* (2004); Baier *et al.* (2007).

[2] We need to extend the mapping to additionally interpret fluent functions and rigid symbols, omitted here for simplicity.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

To account for how knowledge changes after (noise-free) sensing, one defines $w' \sim_z w$, which is to be read as saying "w' and w agree on the sensing for z," as follows:

- if $z = \langle\rangle$, $w' \sim_z w$ for every w'; and
- $w' \sim_{z \cdot a} w$ iff $w' \sim_z w$, $w'[Poss(a), z] = 1$ and $w'[SF(a), z] = w[SF(a), z]$.

This is saying that initially, we would consider all worlds compatible, but after actions, we would need the world w' to agree on the executability of actions performed so far as well as agree on sensing outcomes. The reader might notice that this is clearly a reworking of the successor state axiom for the knowledge fluent in (Scherl and Levesque 2003).

With this, we get a simply account for truth. We define the satisfaction of formulas wrt (with respect to) the triple (e, w, z), and the semantics is defined inductively:

- $e, w, z \models p$ iff p is an atom and $w[p, z] = 1$;
- $e, w, z \models \alpha \land \beta$ iff $e, w, z \models \alpha$ and $e, w, z \models \beta$;
- $e, w, z \models \neg\alpha$ iff $e, w, z \not\models \alpha$;
- $e, w, z \models \forall x\alpha$ iff $e, w, z \models \alpha^x_n$ for all $n \in \mathcal{R}$;
- $e, w, z \models [a]\alpha$ iff $e, w, z \cdot a \models \alpha$;
- $e, w, z \models \Box\alpha$ iff $e, w, z \cdot z' \models \alpha$ for all $z' \in \mathcal{Z}$;
- $e, w, z \models \boldsymbol{K}\alpha$ iff for all $w' \sim_z w$, if $w' \in e$, $e, w', z \models \alpha$; and
- $e, w, z \models \boldsymbol{O}\alpha$ iff for all $w' \sim_z w$, $w' \in e$, iff $e, w', z \models \alpha$.

We write $\Sigma \models \alpha$ (read as "Σ entails α") to mean for every $M = (e, w, \langle\rangle)$, if $M \models \alpha'$ for all $\alpha' \in \Sigma$, then $M \models \alpha$. We write $\models \alpha$ (read as "α is valid") to mean $\{\} \models \alpha$.

Properties. Let us first begin by observing that given a model (e, w, z), we do not require $w \in e$. It is easy to show that if we stipulated the inclusion of the real world in the epistemic state, $\boldsymbol{K}\alpha \supset \alpha$ would be true. That is, suppose $\boldsymbol{K}\alpha$. By the definition above, w is surely compatible with itself after any z, and so α must hold at w. Analogously, properties regarding knowledge can be proven with comparatively simpler arguments in a modal framework, in relation to the classical epistemic situation calculus. Valid properties include:

1. $\Box(\boldsymbol{K}(\alpha) \land \boldsymbol{K}(\alpha \supset \beta) \supset \boldsymbol{K}(\beta))$;
2. $\Box(\boldsymbol{K}(\alpha) \supset \boldsymbol{K}(\boldsymbol{K}(\alpha)))$;
3. $\Box(\neg\boldsymbol{K}(\alpha) \supset \boldsymbol{K}(\neg\boldsymbol{K}(\alpha)))$;
4. $\Box(\forall x.\ \boldsymbol{K}(\alpha) \supset \boldsymbol{K}(\forall x.\ \alpha))$; and
5. $\Box(\exists x.\ \boldsymbol{K}(\alpha) \supset \boldsymbol{K}(\exists x.\ \alpha))$.

Note that such properties hold over all possible action sequences, which explains the presence of the $\Box$ operator on the outside. The first is about the closure of modus ponens within the epistemic modality. The second and third are on positive and negative introspection. The last two reason about quantification outside the epistemic modality, and what that means in terms of the agent's knowledge. For example, item 5 says that if there is some individual n such that the agent knows $Teacher(n)$, it follows that the agent believes $\exists x Teacher(x)$ to be true. This may seem obvious, but note that the property is really saying that the existence of an individual in some possible world implies that such an individual exists in all accessible worlds. It is because there is a fixed domain of discourse that these properties come out true; they are referred to the Barcan formula.

As seen above, the logic $\mathcal{ES}$ allows for a simple definition of the notion of only-knowing in the presence of actions (Levesque 1990), which allows one to capture both the beliefs as well as the non-beliefs of the agent. Using the modal operator $\boldsymbol{O}$ for only-knowing, it can be shown that $\boldsymbol{O}\alpha \models \boldsymbol{K}\beta$ if $\alpha \models \beta$ but $\boldsymbol{O}\alpha \models \neg\boldsymbol{K}\beta$ if $\alpha \not\models \beta$ for any non-modal $\{\alpha, \beta\}$. That is, only-knowing a knowledge base also means knowing everything entailed by that knowledge base. Conversely, it also means not believing everything that is not entailed by the knowledge base. In that sense, $\boldsymbol{K}$ can be seen as an "at least" epistemic operator, and $\boldsymbol{O}$ captures both at least and "at most" knowing. This can be powerful to ensure, for example, that the agent provably does not know protected attributes.

We will now consider the axiomatization of a basic action theory in $\mathcal{ES}$. But before explaining how successor state axioms are written, one might wonder whether a successor state axiom for $\boldsymbol{K}$ is needed, as one would for $Knows$ in the epistemic situation calculus. It turns out because the compatibility of the worlds already accounted for the executability of actions and sensing outcomes in accessible worlds, such an axiom is actually a property of the logic:

$$\models \Box[a]\boldsymbol{K}(\alpha) \equiv (SF(a) \wedge \boldsymbol{K}(SF(a) \supset [a]\alpha)) \ \vee (\neg SF(a) \wedge \boldsymbol{K}(\neg SF(a) \supset [a]\alpha)).$$

(As is usual, free variables are implicitly quantified from the outside.) Thus, what will be known after an action is understood in terms of what was known previously together with the sensing outcome. The example below will further clarify how SF works.

Basic action theories. To axiomatize the domain, we consider the analogue of the basic action theory in the situation calculus (Reiter 2001a). It consists of:

- axioms that describe what is true in the initial states, as well as what is known initially;
- precondition axioms that describe the conditions under which actions are executable using a distinguished predicate $Poss$;
- successor state axioms that describe the conditions under which changes happen to fluents after actions (incorporating Reiter's monotonic solution to the frame problem); and
- sensing axioms that inform the agent about the world using a distinguished predicate SF.

Note that foundational axioms as usually considered in Reiter's variant of the situation calculus (Reiter 2001a) are not needed as the tree-like nature of the situations is baked into the semantics.

Let us consider a simple example of a loan agency set up for the employees of a company. For simplicity, assume actions are always executable: $\Box Poss(a) = true$. Let us also permit a sensing axiom that allows one to look up if an individual is male: $\Box SF(a) \equiv (a = isMale(x) \wedge Male(x)) \vee a \neq isMale(x)$. For simplicity, we assume binary genders, but it is a simple matter of using a predicate such as $Gender(x, y)$ instead to allow individuals x to take on gender y from an arbitrary set.

To now consider successor state axioms, let us suppose having a loan is simply a matter of the manager approving, and unless the manager denies it at some point, the individual continues to hold the loan. For illustration purposes, we will consider a company policy that approves loans for those with high salaries. High salaries are enabled for an "eligible" individual if they are promoted by the manager, and salaries remain high unless they get

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

demoted. Finally, we model eligibility and maleness as a rigid, but this is not necessary, and we can permit actions that updates the gender of individuals in the database. These are formalized as the axioms below, where the left hand side of the equivalence captures the idea that for every sequence of actions, the effect of doing a on a predicate is given by the right hand side of the equivalence.

$$\square[a]hasLoan(x) \equiv a = approve(x) \vee (hasLoan(x) \wedge a \neq deny(x)).$$
$$\square[a]highSalary(x){\equiv}(a{=}promote(x) \wedge Eligible(x)) \vee (highSalary(x) \wedge a{\neq}demote(x)).$$
$$\square[a]Eligible(x) \equiv Eligible(x).$$
$$\square[a]Male(x) \equiv Male(x).$$

We will lump the successor state, precondition and sensing axioms as Σ_{dyn}. The sentences that are true initially will be referred to by Σ_0; however, the agent cannot be expected to know everything that is true, and so let Σ_0' be what is believed initially. It may seem natural to let $\Sigma_0' \subseteq \Sigma_0$, but that is not necessary. The agent might be uncertain about what is true (e.g. Σ_0 might have p but Σ_0' has $p \vee q$ instead).[3] However, for simplicity, we will require that agents at least believe the dynamics works as would the real world. Therefore, we consider entailments wrt the following *background theory*:

$$\Sigma = \Sigma_0 \wedge \Sigma_{dyn} \wedge \boldsymbol{O}(\Sigma_0' \wedge \Sigma_{dyn}). \tag{1}$$

In our example, let us suppose: $\Sigma_0 = \{Male(n_i), \neg Male(n_i'), Eligible(n_i), \neg Eligible(n_i') \mid i \in N\}$ whereas, what is believed by the agent initially is: $\Sigma_0' = \{Eligible(n_i), \neg Eligible(n_i') \mid i \in N\}$ So there are two groups of individuals, n_i and n_i', the first male and the second female, the first considered eligible and the second not considered eligible. All that the agent knows is the eligibility of the individuals. Note that N here is any set, possibly an infinite one, that is, the language allows $N = \mathbb{N}$. For ease of readability, however, we let $N = \{1\}$ in our examples below, and we write n_1 as n and n_1' as n'.[4]

It is worth quickly remarking that many features of the language are omitted here for simplicity. For example, $\mathcal{ES}$ can be extended with second-order variables (Classen and Lakemeyer 2008), which allows one to consider the equivalent of GOLOG programs (Levesque *et al.* 1997). Likewise, notions of probabilistic actions (Bacchus *et al.* 1999), epistemic achievability (Lespérance *et al.* 2000), and causality (Batusov and Soutchanski 2018) in addition to studying program properties (Classen 2018) are interesting dimensions to explore in the fairness context.

Forgetting. In some of the definitions of fairness, we will need to force the setting where information about protected attributes is forgotten. While standard ML approaches propose to do this via column deletion (e.g. remove all entries for the gender

[3] If the agent believes facts that are conflicted by observations about the real world, beliefs may need to be revised (Delgrande and Levesque 2012), a matter we ignore for now. Our theory of knowledge is based on *knowledge expansion* where sensing ensures that the agent is more certain about the world (Scherl and Levesque 2003; Reiter 2001a).

[4] Note that although the language has infinitely many constants, a finite domain can be enforced using domain relativization. For example, let: $\forall x(Individual(x) \equiv x = john \vee \ldots \vee x = jane)$. This declares finitely many individuals. Then instead of saying $\exists x.\ Eligible(x)$, which in general means that any one of the infinitely many constants is eligible, we would write: $\exists x(Individual(x) \wedge Eligible)$, which declares that only one from $\{john, \ldots, jane\}$ is eligible.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

attribute), a richer notion is arguably needed for a first-order knowledge base. We appeal to the notion of forgetting (Lin and Reiter 1994).

Lin and Reiter defined the notion of forgetting, which is adapted to $\mathcal{ES}$ below. They show that while forgetting ground atoms is first-order definable, forgetting relations needs second-order logic. We only focus on the case of atoms, but it would interesting to study how fairness notions are affected when protected attributes are completely absent from a theory.

Suppose S denotes a finite set of ground atoms. We write $\mathcal{M}(S)$ to mean the set of all truth assignments to S. Slightly abusing notation, given a ground atom p, we write $w' \sim_p w$ to mean that w' and w agree on everything initially, except maybe p. That is, for every atom $q \neq p$, $w[q, \langle\rangle] = w'[q, \langle\rangle]$. Next, for every action sequence $z \neq \langle\rangle$ and every atom q', $w[q', z] = w'[q', z]$.

Definition. Given a formula ϕ not mentioning modalities, we say ϕ' is the result of forgetting atom p, denoted $\mathit{Forget}(\phi, p)$, if for any world w, $w \models \phi'$ iff there is a w' such that $w' \models \phi$ and $w \sim_p w'$. Inductively, given a set of atoms $\{p_1, \ldots, p_k\}$, define $\mathit{Forget}(\phi, \{p_1, \ldots, p_k\})$ as $\mathit{Forget}(\mathit{Forget}(\phi, p_1), \ldots, p_k)$.

It is not hard to show that forgetting amounts to setting an atom to true everywhere or setting it false everywhere. In other words:

Proposition. $\mathit{Forget}(\phi, S) \equiv \bigvee_{M \in \mathcal{M}(s)} \phi[M]$, where $\phi[M]$ is equivalent to $\phi \wedge \bigwedge_i (p_i = b_i)$ understood to mean that the proposition p_i is accorded the truth value $b_i \in \{0, 1\}$ by M.

Abusing notation, we extend the notion of forgetting of an atom p for basic action theories and the background theory as follows in applying it solely to what is true/known initially:

- $\mathit{Forget}(\Sigma_0 \wedge \Sigma_{dyn}, p) = \mathit{Forget}(\Sigma_0, p)$; and
- $\mathit{Forget}(\Sigma, p) = \mathit{Forget}(\Sigma_0, p) \wedge \Sigma_{dyn} \wedge \boldsymbol{O}(\mathit{Forget}(\Sigma'_0, p) \wedge \Sigma_{dyn})$.

One of the benefits of lumping the knowledge of the agent as an objective formula in the context of the only-knowing operator is the relatively simple definition of forgetting.

Proposition. Suppose ϕ is non-modal. Suppose p is an atom. For every objective ψ such that $\mathit{Forget}(\phi, p) \models \psi$ it is also the case that $\boldsymbol{O}(\mathit{Forget}(\phi, p)) \models \boldsymbol{K}\psi$.

Because $\boldsymbol{O}\phi \models \boldsymbol{K}\psi$ for every $\{\phi, \psi\}$ provided $\phi \models \psi$, the above statement holds immediately. In so much as we are concerned with a non-modal initial theory and the effects of forgetting, our definition of $\mathit{Forget}(\Sigma, p)$ above (notational abuse notwithstanding) suffices. In contrast, forgetting with arbitrary epistemic logical formulas is far more involved (Zhang and Zhou 2009).

4 Existing notions

As discussed, we will not seek to simply retrofit existing ML notions in a logical language; rather we aim to identify the principles and emphasize the provenance of unfair actions in complex events. Nonetheless, it is useful to revisit a few popular definitions to guide our intuition.

Fairness through unawareness. Fairness through unawareness (FTU) is the simplest definition of fairness; as its name suggests, an algorithm is "fair" if it is unaware of the protected attribute a_p of a particular individual when making a prediction (Kusner *et al.* 2017).

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

Definition. For some set of attributes X any mapping $f : X \to \hat{y}$, where $a_p \notin X$ satisfies fairness through unawareness (Kusner *et al.* 2017). (Assume y denotes the true label.)

This prevents the algorithm learning direct bias on the basis of the protected attribute, but does not prevent indirect bias, which the algorithm can learn by exploiting the relationship between other training variables and the protected attribute (Pedreschi *et al.* 2008; Hardt *et al.* 2016). Moreover, if any of the training attributes are allocated by humans there is the potential for bias to be introduced.

Statistical measures of fairness. Rather than defining fairness in terms of the scope of the training data, much of the existing literature instead assesses whether an algorithm is fair on the basis of a number of statistical criteria that depend on the predictions made by the algorithm (Hardt *et al.* 2016; Kusner *et al.* 2017; Zemel *et al.* 2013). One widely used and simple criterion is demographic parity (DP). In the case that both the predicted outcome and protected attribute a_p are both binary variables, a classifier is said to satisfy predictive parity (Hardt *et al.* 2016) if: $P(\hat{y} = 1|a_p = 1) = P(\hat{y} = 1|a_p = 0)$. By this definition, a classifier is considered fair if it is equally likely to make a positive prediction regardless of the value of the protected attribute a_p.

Fairness and the individual. Another problem with statistical measures is that, provided that the criterion is satisfied, an algorithm will be judged to be fair regardless of the impact on individuals. In view of that, various works have introduced fairness metrics which aim to ensure that individuals are treated fairly, rather than simply considering the statistical impact on the population as a whole (Dwork *et al.* 2011; Kusner *et al.* 2017). Counterfactual fairness (CF), for example, was proposed as a fairness criterion in Kusner *et al.* (2017). The fundamental principle behind this definition of fairness is that the outcome of the algorithm's prediction should not be altered if different individuals within the sample training set were allocated different values for their protected attributes (Kusner *et al.* 2017). This criterion is written in the following form: $P(\hat{y}_{A_p \leftarrow a_p}|A = a, X = x) = P(\hat{y}_{A_p \leftarrow a'_p}|A = a, X = x)\ \forall y, a'$. The notation $\hat{y} \leftarrow_{A_p \leftarrow a_p}$ is understood as "the value of $\hat{y}$ if A_p had taken the value a_p" (Kusner *et al.* 2017).

5 Formalizing Fairness

At the outset, let us note a few salient points about our formalizations of FTU, DP and CF:

1. Because we are not modeling a prediction problem, our definitions below should be seen as being loosely inspired by existing notions rather that faithful reconstructions. In particular, we will look at "fair outcomes" after a sequence of actions. Indeed, debates about problems with the mathematical notions of fairness in single-shot predictions problems are widespread (Dwork *et al.* 2011; Kusner *et al.* 2017; Zafar *et al.* 2017a), leading to recent work on looking at the long-term effects of fairness (Creager *et al.* 2020). However, we are ignoring probabilities in the formalization in current work only to better study the principles behind the above notions – we suspect with a probabilistic epistemic dynamic language (Bacchus *et al.* 1999), the definitions might resemble mainstream notions almost exactly and yet organically use them over actions and programs, which is attractive.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

2. The first-order nature of the language, such as quantification, will allow us to easily differentiate fairness for an individual versus groups. In the mainstream literature, this has to be argued informally, and the intuition grasped meta-linguistically.
3. Because we model the real world in addition the agent's knowledge, we will be able to articulate what needs to be true vs just believed by the agent. In particular, our notion of equity will refer to the real world.
4. De-re vs de-dicto knowledge will mean having versus not having information about protected attributes respectively. Sensing actions can be set up to enable de-re knowledge if need be, but it is easy to see in what follows that de-dicto is preferable.
5. Action sequences can make predicates true, and this will help us think about equity in terms of balancing opportunities across instances of protected attributes (e.g. making some property true so that we achieve gender balance).

Fairness through unawareness. Let us begin with FTU: recall that it requires that the agent does not know the protected attributes of the individuals. To simplify the discussion, let us assume we are concerned with one such attribute $\theta(x)$, say, $Male(x)$, in our examples for concreteness. We might be interested in achieving $hasLoan(x)$ or $highSalary(x)$, for example, either for all x or some individual.

Definition. A sequence $\delta = a_1 \cdots a_k$ implements FTU for ϕ wrt protected attribute $\theta(x)$ iff $\Sigma \models [\delta]\boldsymbol{K}\phi$; and for every $\delta' \leq \delta$: $\Sigma \models [\delta']\neg\exists x(\boldsymbol{K}\theta(x))$.

The attractiveness of a first-order formalism is that in these and other definitions below where we quantify over all individuals, it is immediate to limit the applicability of the conditions wrt specific individuals. Suppose n is such an individual. Then:

Definition. A sequence $\delta = a_1 \cdots a_k$ implements FTU for ϕ wrt attribute $\theta(x)$ for individual n iff (a) $\Sigma \models [\delta]\boldsymbol{K}\phi$; and (b) for every $\delta' \leq \delta$: $\Sigma \models [\delta']\neg\boldsymbol{K}\theta(n)$.

Example. Consider Σ from (1), $Male(x)$ as the protected attribute, and suppose $\delta = approve(n) \cdot approve(n')$. It is clear that δ implements FTU for both the universal $\phi = \forall x hasLoan(x)$ as well as an individual $\phi = hasLoan(n)$. Throughout the history, the agent does not know the gender of the individual.

Before turning to other notions, let us quickly reflect on proxy variables. Recall that in the ML literature, these are variables that indirectly provide information about protected attributes. We might formalize this using entailment:

Definition. Given a protected attribute $\theta(x)$ and theory Σ, let the proxy set $Proxy(\theta(x))$ be the set of predicates $\{\eta_1(x), \ldots \eta_k(x)\}$ such that: $\Sigma \models \forall x(\eta_i(x) \supset \theta(x))$, for $i \in \{1, \ldots, k\}$.

That is, given the axioms in the background theory, $\eta_i(x)$ tells us about $\theta(x)$.

Example. Suppose the agent knows the following sentence: $\forall x(EtonForBoys(x) \supset Male(x))$. Let us assume $EtonForBoys(x)$ is a rigid, like $Male(x)$. Let us also assume that $\boldsymbol{K}(EtonForBoys(n))$. It is clear that having information about this predicate for n would mean the agent can infer that n is male.

The advantage of looking at entailment in our definitions is that we do not need to isolate the proxy set at all, because whatever information we might have the proxy set and its instances, all we really need to check is that $\Sigma \not\models \exists x\boldsymbol{K}\theta(x)$.[5]

[5] With this discussion, we do not mean to insist that analyzing "relevant" predicates for $\theta(x)$ is a pointless endeavor. Rather we only want to point out that regardless of the information available to

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

Demographic parity. Let us now turn to DP. In the probabilistic context, DP is a reference to the proportion of individuals in the domain: say, the proportion of males promoted is the same as the proportion of females promoted. In logical terms, although FTU permitted its definition to apply to both groups and individuals, DP, by definition, is necessarily a quantified constraint. In contrast, CF will stipulate conditions solely on individuals.

Definition. A sequence $\delta = a_1 \cdots a_k$ implements DP for $\phi(x)$ wrt attribute $\theta(x)$ iff: $\Sigma \models [\delta]\boldsymbol{K}((\forall x \theta(x) \supset \phi(x)) \land (\forall x \neg\theta(x) \supset \phi(x)))$.

To reiterate, in probabilistic terms, the proportion of men who are promoted equals the proportion of women who are promoted. In the categorial setting, the agent knows that all men are promoted as well as that all women are promoted.

Example. Consider $\delta = \mathit{approve}(n) \cdot \mathit{approve}(n')$. It implements DP for $\mathit{hasLoan}(x)$ wrt attribute $\mathit{isMale}(x)$.

Note that even though the agent does not know the gender of the individuals, in every possible world, regardless of the gender assigned to an individual n in that world, n has the loan. In other words, all men and all women hold the loan. This is de-dicto knowledge of the genders, and it is sufficient to capture the thrust of DP.

We might be tempted to propose a stronger requirement, stipulating de-re knowledge:

Definition. A sequence $\delta = a_1 \cdots a_k$ implements strong DP for $\phi(x)$ wrt attribute $\theta(x)$ iff: (a)$\Sigma \models [\delta]\boldsymbol{K}((\forall x \theta(x) \supset \phi(x)) \land (\forall x \neg\theta(x) \supset \phi(x)))$; and (b) $\Sigma \models [\delta]\forall x(\boldsymbol{K}\theta(x) \lor \boldsymbol{K}\neg\theta(x))$.

That is, the agent knows whether x is a male or not, for every x.

Example. Consider $\delta = \mathit{isMale}(n) \cdot \mathit{isMale}(n') \cdot \mathit{approve}(n) \cdot \mathit{approve}(n')$. It implements strong DP for $\mathit{hasLoan}(x)$ wrt attribute $\mathit{isMale}(x)$. Of course, by definition, δ also implements DP for $\mathit{hasLoan}(x)$.

FTU-DP. In general, since we do not wish the agent to know the values of protected attributes, vanilla DP is more attractive. Formally, we may impose a FTU-style constraint of not knowing on any fairness definition. For example,

Definition. A sequence $\delta = a_1 \cdots a_k$ implements FTU-DP for $\phi(x)$ wrt attribute $\theta(x)$ iff: (a) $\Sigma \models [\delta]\boldsymbol{K}((\forall x \theta(x) \supset \phi(x)) \land (\forall x \neg\theta(x) \supset \phi(x)))$; and (b) for every $\delta' \leq \delta$: $\Sigma \models [\delta']\neg\exists x \boldsymbol{K}\theta(x)$.

Again, it is worth remarking that mixing and matching constraints is straightforward in a logic, and the semantical apparatus provides us with the tools to study the resulting properties.

Example. The example for de-dicto DP is applicable here too. Consider $\delta = \mathit{approve}(n) \cdot \mathit{approve}(n')$. It implements FTU-DP for $\mathit{hasLoan}(x)$ wrt attribute $\mathit{isMale}(x)$. That is, (a) $\Sigma \not\models \exists x \boldsymbol{K}\theta(x)$; (b) $\Sigma \not\models [\mathit{approve}(n)]\exists x \boldsymbol{K}\theta(x)$; and (c) $\Sigma \not\models [\mathit{approve}(n) \cdot \mathit{approve}(n')]\exists x \boldsymbol{K}\theta(x)$.

Reversing the actions, not surprisingly, $\delta' = \mathit{approve}(n') \cdot \mathit{approve}(n)$ does not affect the matter: δ' also implements FTU-DP. Had the sequence including sensing, a reversal could matter.

the agent, as long as we check that it is actually ignorant about the gender, other relevant predicates may not matter. Of course, a biased agent can enable actions that favors individuals based on such proxy predicates instead, but in that case, such proxy predicates would also need to be included in the protected attribute list.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

One can also consider situations where some knowledge of protected attributes is useful to ensure there is parity but to also account for special circumstances. In this, the protected attribute itself could be "hidden" in a more general class, which is easy enough to do in a relational language.

Example. Suppose we introduce a new predicate for underrepresented groups. We might have, for example: $\forall x(\neg Male(x) \vee \ldots \vee RaceMinority(x) \supset Underrepresented(x))$. This could be coupled with a sensing axiom of the sort: $\square SF(checkU(x)) \equiv Underrepresented(x)$. Add the predicate definition and the sensing axioms to the initial theories and dynamic axioms in Σ respectively. Consider $\delta = checkU(n) \cdot checkU(n') \cdot approve(n) \cdot approve(n')$. Then δ implements strong DP for $hasLoan(x)$ wrt attribute $Underrepresented(x)$. That is, both represented and underrepresented groups have loans.

Equality of opportunity. One problem with DP is that (unless the instance rate of $y = 1$ happens to be the same in both the $a_p = 0$ group and $a_p = 1$ group), the classifier cannot achieve 100% classification accuracy and satisfy the fairness criterion simultaneously (Hardt *et al.* 2016). Also, there are scenarios where this definition is completely inappropriate because the instance rate of $y = 1$ differs so starkly between different demographic groups. Finally, there are also concerns that statistical parity measures fail to account for fair treatment of individuals (Dwork *et al.* 2011). Nonetheless it is often regarded as the most appropriate statistical definition when an algorithm is trained on historical data (Zafar *et al.* 2017b; Zemel *et al.* 2013).

A modification of demographic parity is "equality of opportunity" (EO). By this definition, a classifier is considered fair if, among those individuals who meet the positive criterion, the instance rate of correct prediction is identical, regardless of the value of the protected attribute (Hardt *et al.* 2016). This condition can be expressed as (Hardt *et al.* 2016): $P(y = 1|a_p = a, \hat{y} = 1) = P(y = 1|a_p = a', \hat{y} = 1) \ \forall\, a, a'$. In (Hardt *et al.* 2016), it is pointed out that a classifier can simultaneously satisfy equality of opportunity and achieve perfect prediction whereby $\hat{y} = y$ (prediction=true label) in all cases.

In the logical setting, this can be seen as a matter of only looking at individuals that satisfy a criterion, such as being eligible for promotion or not being too old to run for office.

Definition. A sequence δ implements EO for $\phi(x)$ wrt attribute $\theta(x)$ and criterion $\eta(x)$ iff:

$$\Sigma \models [\delta]\boldsymbol{K}((\forall x(\eta(x) \wedge \theta(x)) \supset \phi(x)) \wedge (\forall x \neg(\eta(x) \wedge \theta(x)) \supset \phi(x))).$$

Example. Consider $\delta = promote(n) \cdot promote(n')$, let $\phi(x) = highSalary(x)$ and the criterion $\eta(x) = Eligible(x)$. Although the promote action for n' does not lead her to obtain a high salary, because we condition the definition only for eligible individuals, δ does indeed implement EO. Note again that the agent does not know the gender for n', but in every possible world, regardless of the gender n' is assigned, n' is known to be ineligible. In contrast, n is eligible and δ leads to n having a high salary. That is, every eligible male now has high salary, and every eligible female also has high salary. (It just so happens there are no eligible females, but we will come to that.)

In general, the equality of opportunity criterion might well be better applied in instances where there is a known underlying discrepancy in positive outcomes between two different groups, and this discrepancy is regarded as permissible. However, as we might

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

observe in our background theory, there is systematic bias in that no women is considered eligible.

Counterfactual fairness. Let us now turn to CF. The existing definition forces us to consider a "counterfactual world" where the protected attribute values are reversed, and ensure that the action sequence still achieves the goal.

Definition. A sequence $\delta = a_1 \cdots a_k$ implements CF for ϕ wrt attribute $\theta(x)$ for individual n iff:

- $\Sigma \models (\theta(n) = b)$ for $b \in \{0, 1\}$ and $\Sigma \models [\delta]\boldsymbol{K}\phi$; and
- $\mathit{Forget}(\Sigma, \theta(n)) \wedge (\theta(n) \neq b) \models [\delta]\boldsymbol{K}\phi$.

Example. Let us consider the case of loan approvals. Consider the individual n and the action $\delta = \mathit{approve}(n)$. Let $\phi = \mathit{hasLoan}(n)$, and the protected attribute $\mathit{Male}(x)$. Clearly $\Sigma \models \mathit{Male}(n)$, and indeed $\Sigma \models [\delta]\mathit{hasLoan}(n)$. If we consider Σ' where the gender for n is swapped, it is still the case that $\Sigma' \models [\delta]\mathit{hasLoan}(n)$. Thus δ implements CF for $\mathit{hasLoan}(n)$ wrt $\mathit{Male}(n)$.

The definition of CF is well-intentioned, but does not quite capture properties that might enable equity. Indeed, there is a gender imbalance in the theory, in the sense that only the male employee is eligible for promotions and the female employee can never become eligible. Yet CF does not quite capture this. Let us revisit the example with getting high salaries:

Example. Consider $\delta = \mathit{promote}(n)$ for property $\mathit{highSalary}(n)$ wrt attribute $\mathit{Male}(n)$. It is clear that δ implements CF because the gender is irrelevant given that n is eligible. However, given $\delta' = \mathit{promote}(n')$, we see that δ' does not implement CF for $\mathit{highSalary}(n')$ wrt $\mathit{Male}(n')$. Because n' is not eligible, $\mathit{highSalary}(n')$ does not become true after the promotion.

Equity. Among the many growing criticisms about formal definitions of fairness is that notions such as CF fail to capture systemic injustices and imbalances. We do not suggest that formal languages would address such criticisms, but they provide an opportunity to study desirable augmentations to the initial knowledge or action theory.

Rather than propose a new definition, let us take inspiration from DP, which seems fairly reasonable except that it is the context of what the agent knows. Keeping in mind a desirable "positive" property such as $\mathit{Eligible}(x)$, let us consider DP but at the world-level:

Definition. Given a theory Σ, protected attribute $\theta(x)$, positive property $\eta(x)$, where x is the individual, define *strong equity*: $\Sigma \models \forall x(\theta(x) \supset \eta(x)) \wedge \forall x(\neg\theta(x) \supset \eta(x))$.

In general, it may not be feasible to ensure that properties hold for all instances of both genders. For example, there may be only a handful of C-level executives, and we may wish that there are executives of both genders.

Definition. Given a theory Σ, protected attribute $\theta(x)$, positive property $\eta(x)$, where x is the individual, define *weak equity*: $\Sigma \models \exists x(\theta(x) \wedge \eta(x)) \wedge \exists x(\neg\theta(x) \wedge \eta(x))$. It is implicitly assumed that the set of positive and negative instances for $\theta(x)$ is non-empty: that is, assume the integrity constraint: $\Sigma \models \exists x, y(\theta(x) \wedge \neg\theta(y))$.

We assume weak equity and focus on FTU below. The definitions could be extended to strong equity or other fairness notions depending on the modeling requirements.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

Definition. A sequence $\delta = a_1 \cdots a_k$ implements equitable FTU for ϕ wrt protected attribute $\theta(x)$ and property $\eta(x)$ iff (a) either weak equity holds in Σ and δ implements FTU; or (b) δ implements equitable FTU for ϕ wrt $\theta(x)$ and $\eta(x)$ for the updated theory $\mathit{Forget}(\Sigma, S)$, where $S = \{\eta(n_i) \mid i \in N\}$.

Note that we are assuming that N is finite here because we have only defined forgetting wrt finitely many atoms. Otherwise, we would need a second-order definition.

Example. Consider $\delta = \mathit{promote}(n) \cdot \mathit{promote}(n')$ for goal $\phi = \forall x(\mathit{highSalary}(x))$ wrt protected attribute $\mathit{Male}(x)$ and property $\mathit{Eligible}(x)$. It is clear that weak equity does not hold for Σ because there is a female who is not eligible. In this case, consider $\Sigma' = \mathit{Forget}(\Sigma, S)$ where $S = \{\mathit{Eligible}(n), \mathit{Eligible}(n')\}$. And with that, Σ' also does not mention that n is eligible, so the promotion actions does not lead to anyone having high salaries. So δ does not enable knowledge of ϕ.

Example. Let us consider Σ' that is like Σ except that $\mathit{Eligible}(x)$ is not rigid, and can be affected using the action $\mathit{make}(x)$: $\Box[a]\mathit{Eligible}(x) \equiv \mathit{Eligible}(x) \vee (a = \mathit{make}(x))$. That is, either an individual is eligible already or the manager makes them. Of course, $\delta = \mathit{promote}(n) \cdot \mathit{promote}(n')$ from above still does not implement equitable FTU, because we have not considered any actions yet to make individuals eligible. However, consider $\delta' = \mathit{make}(n) \cdot \mathit{make}(n') \cdot \mathit{promote}(n) \cdot \mathit{promote}(n')$. Because Σ does not satisfy weak equity, we turn to the second condition of the definition. On forgetting, no one is eligible in the updated theory, but the first two actions in δ' makes both n and n' eligible, after which, they are both promoted. So δ' enables knowledge of $\forall x(\mathit{highSalary}(x))$. Thus, the actions have made clear that eligibility is the first step in achieving gender balance, after which promotions guarantee that there are individuals of both genders with high salaries.

6 Conclusions

In this paper, we looked into notions of fairness from the machine learning literature, and inspired by these, we attempted a formalization in an epistemic logic. Although we limited ourselves to categorical knowledge and noise-free observations, we enrich the literature by considering actions. Consequently we looked into three notions: fairness through unawareness, demographic parity and counterfactual fairness, but then expanded these notions to also tackle equality of opportunity as well as equity. We were also able to mix and match constraints, showing the advantage of a logical approach, where one can formally study the properties of (combinations of) definitions. Using a simple basic action theory we were nonetheless able to explore these notions using action sequences.

As mentioned earlier, this is only a first step and as argued in works such as Pagnucco *et al.* (2021); Dehghani *et al.* (2008); Halpern and Kleiman-Weiner (2018) there is much promise in looking at ethical AI using rich logics. In fact, we did not aim to necessarily faithfully reconstruct existing ML notions in this paper but rather study underlying principles. This is primarily because we are not focusing on single-shot prediction problems but how actions, plans and programs might implement fairness and de-biasing. The fact that fairness was defined in terms of actions making knowledge of the goal true, exactly as one would in planning (Levesque 1996), is no accident.

State-of-the-art analysis in fairness is now primarily based on false positives and false negatives (Verma and Rubin 2018). So we think as the next step, a probabilistic language

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

such as Bacchus *et al.* (1999) could bring our notions closer to mainstream definitions, but now in the presence of actions. In the long term, the goal is to logically capture bias in the presence of actions as well as repeated harms caused by systemic biases (Creager *et al.* 2020). Moreover, the use of logics not only serve notions such as verification and correctness, but as we argue, could also provide a richer landscape for exploring ethical systems, in the presence of background knowledge and context. This would enable the use of formal tools (model theory, proof strategies and reasoning algorithms) to study the long-term impact of bias while ensuring fair outcomes throughout the operational life of autonomous agents embedded in complex socio-technical applications.

Of course, a logical study such as ours perhaps has the downside that the language of the paper is best appreciated by researchers in knowledge representation, and not immediately accessible to a mainstream machine learning audience. But on the other hand, there is considerable criticism geared at single-shot prediction models for not building in sufficient context and commonsense. In that regard, operationalizing a system that permits a declaration of the assumptions and knowledge of the agents and their actions might be exactly "what the doctor ordered." See also efforts in causal modeling (Chockler and Halpern 2004) that are close in spirit.

References

Alexander, L. and Moore, M. 2016. Deontological ethics. In *The Stanford Encyclopedia of Philosophy*.

Anderson, M. and Anderson, S. L. 2014. Geneth: A general ethical dilemma analyzer. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 253–261.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. 2016. "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.".

Aziz, H., Bouveret, S., Caragiannis, I., Giagkousi, I. and Lang, J. 2018. Knowledge, fairness, and social constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Bacchus, F., Grove, A. J., Halpern, J. Y. and Koller, D. 1996. From statistical knowledge bases to degrees of belief. *Artificial intelligence*, *87*, 1–2, 75–143.

Bacchus, F., Halpern, J. Y. and Levesque, H. J. 1999. Reasoning about noisy sensors and effectors in the situation calculus. *Artificial Intelligence*, *111*, 1–2, 171 – 208.

Baier, J. A., Fritz, C. and McIlraith, S. A. 2007. Exploiting procedural domain control knowledge in state-of-the-art planners. In *Proceedings of ICAPS*, 26–33.

Baral, C., Bolander, T., van Ditmarsch, H. and McIlrath, S. 2017. Epistemic planning (Dagstuhl seminar 17231). *Dagstuhl Reports*, *7*, 6, 1–47.

Batusov, V. and Soutchanski, M. 2018. Situation calculus semantics for actual causality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

Belle, V. 2022. Analyzing generalized planning under nondeterminism. *Artificial Intelligence*, 103696.

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

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z. and 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*, 9–16.

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

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

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

Classen, J. 2018. Symbolic verification of Golog programs with first-order BDDs. In *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)*, M. Thielscher, F. Toni and F. Wolter, Eds., AAAI Press, 524–528.

Classen, J. and Delgrande, J. 2020. Dyadic obligations over complex actions as deontic constraints in the situation calculus. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, vol. 17, 253–263.

Classen, J., Engelmann, V., Lakemeyer, G. and Röger, G. 2008. Integrating Golog and planning: An empirical evaluation. In *NMR Workshop*, 10–18.

Classen, J. and Lakemeyer, G. 2008. A logic for non-terminating golog programs. In *KR*, 589–599.

Conway, P. and Gawronski, B. 2013. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of Personality and Social Psychology*, *104*, 2, 216.

Creager, E., Madras, D., Pitassi, T. and Zemel, R. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning* 2020, pp. 2185–2195. PMLR.

Czelakowski, J. 1997. Action and deontology. In *Logic, Action and Cognition*. Springer, 47–87.

Dehghani, M., Tomai, E., Forbus, K. D. and Klenk, M. An integrated reasoning approach to moral decision-making. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence* 2008, pp. 1280–1286.

Delgrande, J. P. and Levesque, H. J. Belief revision with sensing and fallible actions. In *Proc. KR* 2012.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. 2011. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*.

Farnadi, G., Babaki, B. and Getoor, L. 2018. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 108–114.

Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., et al.. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 69–78.

Georgeff, M., Pell, B., Pollack, M., Tambe, M. and Wooldridge, M. 1998. The belief-desire-intention model of agency. In *International Workshop on Agent Theories, Architectures, and Languages*. Springer, 1–10.

Halpern, J. Y. and Kleiman-Weiner, M. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 1853–1860.

Halpern, J. Y. and Moses, Y. 2014. A procedural characterization of solution concepts in games. *JAIR*, *49*, 143–170.

Halpern, J. Y., Pass, R. and Raman, V. 2009. An epistemic characterization of zero knowledge. In *TARK*, 156–165.

Hardt, M., Price, E. and Srebro, N. 2016. Equality of opportunity in supervised learning. In *International Conference on Neural Information Processing Systems*.

Hooker, J. N. and Kim, T. W. N. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 130–136.

Ignatiev, A., Cooper, M. C., Siala, M., Hebrard, E. and Marques-Silva, J. 2020. Towards formal fairness in machine learning. In *Principles and Practice of Constraint Programming: 26th International Conference, CP 2020, Louvain-la-Neuve, Belgium, September 7–11, 2020, Proceedings 26*. Springer, 846–867.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

Kawamoto, Y. 2019. Towards logical specification of statistical machine learning. In *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings*. Springer, 293–311.

Khandani, A., Kim, J. and Lo, A. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance 34*.

Kusner, M., Loftus, J., Russel, C. and Silva, R. 2017. Counterfactual fairness. *Neural Information Processing Systems*.

Lakemeyer, G. and Levesque, H. J. 2004. Situations, Si! situation terms, No! In *Proceedings of KR* 2004, 516–526.

Lakemeyer, G. and Levesque, H. J. 2007. Cognitive robotics. In *Handbook of Knowledge Representation*. Elsevier, 869–886.

Lakemeyer, G. and Levesque, H. J. 2011. A semantic characterization of a useful fragment of the situation calculus with knowledge. *Artificial Intelligence*, *175*, 142–164.

Lee, J. and Palla, R. 2012. Reformulating the situation calculus and the event calculus in the general theory of stable models and in answer set programming. *Journal of Artificial Intelligence Research*, *43*, 571–620.

Lespérance, Y., Levesque, H. J., Lin, F. and Scherl, R. B. 2000. Ability and knowing how in the situation calculus. *Studia Logica*, *66*, 1, 165–186.

Levesque, H., Reiter, R., Lespérance, Y., Lin, F. and Scherl, R. 1997. Golog: A logic programming language for dynamic domains. *Journal of Logic Programming*, *31*, 59–84.

Levesque, H. J. 1990. All I know: A study in autoepistemic logic. *Artificial Intelligence*, *42*, 2–3, 263–309.

Levesque, H. J. 1996. What is planning in the presence of sensing? In *Proceedings of AAAI / IAAI*, 1139–1146.

Lin, F. and Reiter, R. 1994. Forget it. In *Working Notes of AAAI Fall Symposium on Relevance*, 154–159.

Lindner, F., Bentzen, M. M. and Nebel, B. 2017. The hera approach to morally competent robots. In *Proceedings of the IEEE/RSJ Intelligent Robots and Systems*, 6991–6997.

Liu, X. and Lorini, E. 2022. A logic of "black box" classifier systems. In *Logic, Language, Information, and Computation: 28th International Workshop, WoLLIC 2022, Iaşi, Romania, September 20–23, 2022, Proceedings*. Springer, 158–174

McIlraith, S. A. and Son, T. C. 2002. Adapting golog for composition of semantic web services. In *KR*, 482–496.

Muggleton, S., De Raedt, L., Poole, D., Bratko, I., Flach, P., Inoue, K. and Srinivasan, A. 2012. Ilp turns 20. *Machine Learning*, *86*, 1, 3–23.

Muise, C., Belle, V., Felli, P., McIlraith, S., Miller, T., Pearce, A. and Sonenberg, L. 2015. Planning over multi-agent epistemic states: A classical planning approach. In *Proceedings of AAAI*.

Pagnucco, M., Rajaratnam, D., Limarga, R., Nayak, A. and Song, Y. 2021. Epistemic reasoning for machine ethics with situation calculus. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 814–821.

Pedreschi, D., Ruggieri, S. and Turini, F. 2008. Discrimination aware data mining. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

Reiter, R. 2001a. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press.

Reiter, R. 2001b. On knowledge-based programming with sensing in the situation calculus. *ACM Transactions on Computational Logic*, *2*b, 4, 433–457.

https://doi.org/10.1017/S1471068423000157 Published online by Cambridge University Press

Sardina, S., De Giacomo, G., Lespérance, Y. and Levesque, H. J. 2004. On the semantics of deliberation in indigolog—from theory to implementation. *Annals of Mathematics and Artificial Intelligence*, *41*, 2–4, 259–299.

Sardina, S. and Lespérance, Y. 2010. Golog speaks the BDI language. In *Programming Multi-Agent Systems*, vol. 5919. LNCS. Springer Berlin Heidelberg, 82–99.

Scherl, R. B. and Levesque, H. J. 2003. Knowledge, action, and the frame problem. *Artificial Intelligence*, *144*, 1-2, 1–39.

Verma, S. and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*. IEEE, 1–7.

Wang, K. and Zhang, Y. 2005. Nested epistemic logic programs. In *International Conference on Logic Programming and Nonmonotonic Reasoning*. Springer, 279–290.

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

Zafar, M. B., Valera, I., Rodriguez, M. G. and Gummadi, K. P. 2017b. International conference on artificial intelligence and statistics. In *Fairness Constraints: Mechanisms for Fair Classification*.

Zemel, R., Wu, Y., Swersky, K. and Pitassi, T. 2013. Learning fair representations. In *International Conference on Machine Learning*.

Zhang, Y. and Zhou, Y. 2009. Knowledge forgetting: Properties and applications. *Artificial Intelligence*, *173*, 16–17, 1525–1537.