Modeling Bias in Machine Learning with Description Logics and Epistemic Modalities (Extended Abstract)

Mattia Petrolo[1], Ekaterina Kubyshkina[2]

[1]LASIGE Computer Science and Engineering Research Centre, University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal
[2]Department of Philosophy "Piero Martinetti", University of Milan, via Festa del Perdono, 7 - 20122, Milan, Italy

Abstract

The identification and mitigation of bias in machine learning systems requires a precise formal representation of the norms against which inferences are evaluated. Formal ontologies provide a natural foundation for this task. In this work, we develop a Description Logic–based general formalization of the structure underlying several types of bias, with explicit attention to the epistemic states of agents. First, we employ the Description Logic $\mathcal{ALC}$ to axiomatize the ontological structure that encodes the normative conditions under which an inference counts as biased or unbiased. Second, we provide a translation of the $\mathcal{ALC}$ formalization into a first-order modal framework enriched with necessity and knowledge operators. This step allows us to integrate ontological constraints with representations of agents' epistemic attitudes, thereby connecting DL-based knowledge representation with epistemic reasoning. The combined framework makes it possible to characterize formally the epistemic conditions under which agents may generate biased inferences and to distinguish them from configurations in which bias cannot arise.

Keywords

Bias in Machine Learning, Knowledge Representation, $\mathcal{ALC}$, Epistemic Logic

1. Introduction

In recent years, formal ontologies have taken on a central role in analyzing salient features of Machine Learning (ML) systems (see, e.g., [1], [2], [3]). In particular, Ferrario et al. [4] examine bias through the lens of applied ontology, modeling it in terms of ontological descriptions in DOLCE. Their study explores how systematic deviations emerge from interactions among observations, acquired knowledge, and the classifications produced by ML systems. They argue that uncovering and mitigating bias first requires identifying and representing it, a task for which formal ontologies are especially well suited because they offer conceptual clarity and a shared vocabulary across the ML pipeline. The authors maintain that bias evaluation is always relative to a reference norm, whether statistical, social, ethical, or legal, since the same phenomenon may be considered biased under one standard but not another. Accordingly, they define biases as concepts classifying inferences grounded in specific observations, prior knowledge, and resulting classifications, and distinguish four types of bias —sampling, specification, inductive, and deployment—based on which element in this inferential process deviates from the relevant norm.

Ferrario et al.'s framework is epistemologically focused and, as such, agent-based. Building on their approach, in this work, we place a more explicit emphasis on the epistemic states of agents. To this end, we adopt Kripke semantics, where accessibility relations are epistemic and possible worlds represent epistemic possibilities over ontological states. This setting enables us to model agents and their reasoning about alternative ontological configurations. Our goal is to formally characterize the conditions under which an agent's reasoning may be affected by outputs that are assessed as biased. Our main contribution is a unified framework that integrates ontological constraints with epistemic representations of agents interacting with ML systems. Building on a norm-theoretic conception of bias, according to which bias consists in a systematic departure from a relevant statistical or social norm (see [5]), we analyze such conditions through two interconnected dimensions:

(i) normative principles that relate ontologies to admissible or expected outputs, and

DL 2026: 39th International Workshop on Description Logics, July 17–19, 2026, Lisbon, Portugal
mpetrolo@fc.ul.pt (M. Petrolo); ekaterina.kubyshkina@unimi.it (E. Kubyshkina)

(ii) the epistemic attitudes of agents toward those ontologies and outputs.

On this basis, we develop a formal framework that supports systematic reasoning about different types of bias and allows us to identify specific reasons for which an agent may – or may not – be subject to a given form of bias. From a technical standpoint, we embed ontological descriptions formulated in $\mathcal{ALC}$ within an epistemic framework. More specifically, we translate $\mathcal{ALC}$ into first-order logic and extend the resulting language with two modal operators, knowledge and necessity, interpreted over multi-agent first-order Kripke models. This extension allows us to combine DL-based ontological constraints with the explicit representation of agents' epistemic states. Within this enriched framework, the four bias types introduced by Ferrario et al. can be reformulated in explicitly epistemic terms.

For the purposes of this extended abstract, we focus on a general formalization of the structure underlying sampling, specification, inductive, and deployment bias, illustrating the expressive adequacy and methodological benefits of the proposed approach. The concrete formalization of each bias type can be systematically derived within the same setting. A detailed formal development of these cases, together with a systematic comparison among them, is deferred to a longer version of the paper.

2. A formal analysis

The starting point of our analysis is an extension of the syntax of $\mathcal{ALC}$ (see, e.g., [6, pp. 10 - 139]) with the modal operators $\Box$ and K_a.[1] We denote this extended framework $\mathcal{ALC}^+$ and define its language as follows, where A stands for atomic concepts:

$$C ::= A \mid \top \mid \bot \mid C \sqcap C \mid C \sqcup C \mid \neg C \mid \forall r.C \mid \exists r.C$$

$$\psi ::= C(a) \mid R(a,a) \mid C \sqsubseteq C \mid \Box\psi \mid K_a\psi$$

Concepts and relations among them are interpreted as usual. The operators $\Box$ and K_a are applicable to three types of expression: concept assertions $C(a)$, role assertions $R(a,b)$, and inclusion statements $C \sqsubseteq D$. We refer to such expressions as ψ. Expressions of the form ψ, $\Box\psi$, $K_a\psi$ constitute the formulas of $\mathcal{ALC}^+$. Formulas of a form $\Box\psi$ and $K_a\psi$ denote that ψ is necessary and ψ is known by agent a, respectively. We now define the Kripke frames and models used to interpret $\mathcal{L}^{\mathcal{ALC}+}$.

Definition 1. *A frame is a pair $\mathcal{F} = \langle W, R^A \rangle$ where W is a non-empty set of possible worlds, R^A is a function, yielding for every agent $a \in A$ an accessibility relation $R^A(a) \subseteq W \times W$ (we will write R_a rather then $R^A(a)$). A model is a tuple $\mathcal{M} = \langle \mathcal{F}, \mathcal{I} \rangle$ where $\mathcal{F}$ is a frame and $\mathcal{I}$ is a function that associates to each $w \in W$ a model of $\mathcal{ALC}$.*

A concept C is satisfied in $\mathcal{M}$ if there is w in $\mathcal{F}$ such that $C^{\mathcal{I}_w} \neq \emptyset$, and that C is satisfiable if there is a model in which it is satisfied. The satisfaction of a formula ψ in w of $\mathcal{M}$ ($\mathcal{M}, w \models \psi$), is defined as follows:

- $\mathcal{M}, w \models C \sqsubseteq D$ *iff* $C^{\mathcal{I}_w} \subseteq D^{\mathcal{I}_w}$,
- $\mathcal{M}, w \models r(a,b)$ *iff* $(a^{\mathcal{I}_w}, b^{\mathcal{I}_w}) \in r^{\mathcal{I}_w}$,
- $\mathcal{M}, w \models \phi \wedge \psi$ *iff* $\mathcal{M}, w \models \phi$ *and* $\mathcal{M}, w \models \psi$,
- $\mathcal{M}, w \models C(a)$ *iff* $a^{\mathcal{I}_w} \in C^{\mathcal{I}_w}$,
- $\mathcal{M}, w \models \neg\phi$ *iff* $\mathcal{M}, w \not\models \psi$,
- $\mathcal{M}, w \models \Box\phi$ *iff for all* $w' \in W$, $\mathcal{M}, w' \models \phi$,

[1] While [7] introduce an epistemic modality K that applies to roles and concepts, our aim is instead to define modalities that apply to assertions, rather than to roles or concepts themselves. The same observation applies to more recent work in Epistemic Description Logics, such as [8], who adapt Donini's definition of the epistemic operator K. Our approach, however, is closer to other proposals in Epistemic Description Logics, such as [9], where the operator K is introduced directly at the first-order level. Nevertheless, the two approaches differ in at least two important respects. First, although the semantic definition of the K operator in first-order logic is essentially the same, in our framework it is not intended to represent the knowledge encoded in a knowledge base. Rather, it models the knowledge of agents interacting with an ML system. Consequently, our interpretation of K concerns epistemic possibilities in the standard sense of epistemic logic. Second, from a technical perspective, our framework provides a uniform treatment of both the K and $\Box$ operators.

- $\mathcal{M}, w \models K_a\phi$ *iff for all* w' *if* wR_aw' *then* $\mathcal{M}, w' \models \phi$.

We say that ϕ *is satisfied in* $\mathcal{M}$ *if there is* $w \in W$ *such that* $\mathcal{M}, w \models \phi$*, and that* ϕ *is satisfiable if it is satisfied in some model;* ϕ *is valid in* $\mathcal{M}$ *(*$\mathcal{M} \models \phi$*) , if it is satisfied in all* w *of* $\mathcal{M}$*; and* ϕ *is valid on* $\mathcal{F}$ *(*$\mathcal{F} \models \phi$*) if, for all* $\mathcal{M}$ *based on* $\mathcal{F}$*,* ϕ *is valid in* $\mathcal{M}$.

In what follows, we use $\mathcal{O}$ to denote the conjunction of all axioms of a given ontology defined in $\mathcal{ALC}^+$ terms. Note that $\mathcal{O}$ is not necessarily a formula. As such, modal operators ($\Box$ and K_a) cannot be applied directly to $\mathcal{O}$. In order to meaningfully express modal statements about ontologies, we therefore introduce a first-order modal translation of $\mathcal{ALC}^+$, which allows modalities to range over the translated axioms.

Definition 2. *A first-order modal model is a tuple* $\mathcal{M}^{FO} = \langle W, R^A, D, I\rangle$*, where* W *is a non-empty set of worlds,* R^A *is a function, yielding for every agent* $a \in A$ *an accessibility relation* $R^A(a) \subseteq W \times W$ *(we will often write simply* R_a *instead of* $R^A(a)$*),* D *is a non-empty set, called domain,* I *is a function that to each* n*-place predicate symbol* P *in* w *assigns a subset of* D^n *(*$I(P, w) \subseteq D^n$*).*

A *substitution* is any function that to each variable assigns an element of D ($\sigma : Var \rightarrow D$). A substitution σ' is said to be an x*-variant* of σ if $\sigma(y) = \sigma'(y)$ for all variable y except possibly x, denoted $\sigma \sim_x \sigma'$. Truth is defined in a state relative to a substitution.

Definition 3 (First-order models. Truth conditions). *Let* $\mathcal{M}^{FO} = \langle W, R^A, D, I\rangle$*. The relation* $\mathcal{M}, w \models_\sigma \phi$ *is defined by induction, where* w *is a world,* σ *is an assignment, and* ϕ *is a formula of first-order modal logic.*

- $\mathcal{M}^{FO}, w \models_\sigma P(x_1, ..., x_n)$ *iff* $\langle \sigma(x_1), ..., \sigma(x_n)\rangle \in I(P, w)$ *for each* n*-place predicate symbol* P
- $\mathcal{M}^{FO}, w \models_\sigma \phi \wedge \psi$ *iff* $\mathcal{M}^{FO}, w \models_\sigma \phi$ *and* $\mathcal{M}^{FO}, w \models_\sigma \psi$
- $\mathcal{M}^{FO}, w \models_\sigma \neg\phi$ *iff* $\mathcal{M}^{FO}, w \not\models_\sigma \phi$
- $\mathcal{M}^{FO}, w \models_\sigma \forall x\phi$ *iff for any* $\sigma \sim_x \sigma'$*,* $\mathcal{M}^{FO}, w \models_{\sigma'} \phi$
- $\mathcal{M}^{FO}, w \models_\sigma \Box\phi$ *iff for all* w'*,* $\mathcal{M}^{FO}, w' \models_\sigma \phi$
- $\mathcal{M}^{FO}, w \models_\sigma K_a\phi$ *iff for all* w'*, if* wR_aw' *then* $\mathcal{M}^{FO}, w' \models_\sigma \phi$

A formula ϕ *is said to be true in the world* w *if* $\mathcal{M}^{FO}, w \models_\sigma \phi$*; otherwise, it is said to be false at* w*. If* $\mathcal{M}^{FO}, w \models_\sigma \phi$ *for every world* w*,* $\mathcal{M}^{FO} \models_\sigma \phi$*. If* $\mathcal{M}^{FO} \models_\sigma \phi$ *for any assignment* σ*,* $\mathcal{M}^{FO} \models \phi$.

This semantics characterizes the following system.

Definition 4 (System FO).

All classical tautologies	$(taut)$
S5 axioms for $\Box$	$(S5_\Box)$
K axioms for K_i	(K_{K_i})
$\Box\phi \rightarrow K_i\Box\phi$	$(universal\ K)$
$\bigcirc\phi \rightarrow \Box \bigcirc \phi$ *for* $\bigcirc = K_i, \neg K_i\neg$	$(\bigcirc - uni)$
Modus Ponens	(MP)
Necessitation for $\bigcirc = \Box, K_i$	$(N_\bigcirc)$
$\forall x\phi(x) \rightarrow \phi[y/x]$	$(\forall)$
$\forall x \bigcirc \phi(x) \rightarrow \bigcirc\forall x\phi(x)$ *for* $\bigcirc = \Box, K_i$	(BF)

Theorem 1. *System FO is sound and complete with respect to the semantics of Definition 3.*

We are now in a position to define a translation from $\mathcal{ALC}^+$ into FO. This translation extends the one presented in [6, p. 45] with the following cases for $\Box\phi$ and $K_a\phi$: $\pi_x(\Box\phi) = \Box\pi_x(\phi)$; $\pi_x(K_a\phi) = K_a\pi_x(\phi)$; $\pi_y(\Box\phi) = \Box\pi_y(\phi)$; $\pi_y(K_a\phi) = K_a\pi_y(\phi)$. The fact that this translation preserves the semantics can be established by a straightforward induction.

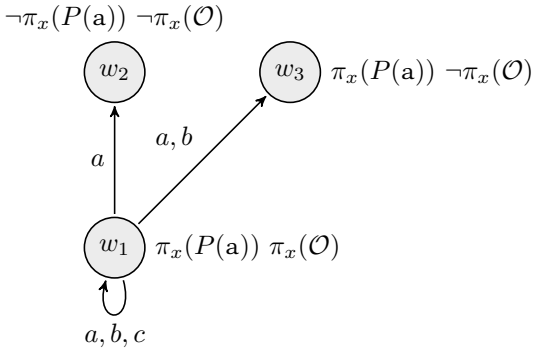

Figure 1: Model $\mathcal{M}^{bias}$

3. A general scheme for bias

Having established a sound and complete logic for reasoning about agents' attitudes (FO), we are now in a position to represent a common root underlying the biases identified in [4]. Informally, this root can be characterized as a deviation of agents' beliefs from an appropriate normative standard, which manifests across different kinds of bias. From this perspective, two main aspects must be accounted for: the relevant normative standard and the epistemic states of the agents.

Consider an ML system that can be described in terms of an ontology $\mathcal{O}$. Suppose that a normative constraint requires that this system, given $\mathcal{O}$, produce a specific output $P(\mathrm{a})$. This requirement can be formalized by stipulating that $\mathcal{O}$ implies $P(\mathrm{a})$ across all epistemically possible worlds for every agent. Now consider an agent a who takes it to be epistemically possible that, given the ontology $\mathcal{O}$, the output $P(\mathrm{a})$ is not produced. Semantically, this situation can be represented within a first-order modal model as follows, where $\otimes$ denotes an operator capturing a general form of bias.

Definition 5. *Given $\mathcal{M}^{FO}$ and $w, w', w'' \in W$ of $\mathcal{M}^{FO}$:*
$\mathcal{M}^{FO}, w \models_\sigma \otimes_a(\pi_x(\mathcal{O}), \pi_x(P(\mathrm{a})))$ iff for all $w' \in W$, $\mathcal{M}^{FO}, w' \models_\sigma \pi_x(\mathcal{O})$ implies $\mathcal{M}^{FO}, w' \models_\sigma \pi_x(P(\mathrm{a}))$; and there exists at least one w'' such that wR_aw'' and if $\mathcal{M}^{FO}, w'' \models_\sigma \pi_x(\mathcal{O})$ then $\mathcal{M}^{FO}, w'' \not\models_\sigma \pi_x(P(\mathrm{a}))$.

With this definition, $\pi_x(\mathcal{O})$ and $\pi_x(P(\mathrm{a}))$ denote the translations of the ontology $\mathcal{O}$ and of the output $P(\mathrm{a})$ into the first-order language. The expression $\otimes_a(\pi_x(\mathcal{O}), \pi_x(P(\mathrm{a}))$ indicates that agent a is subject to a bias when evaluating the output $P(\mathrm{a})$ in the context of an ontology $\mathcal{O}$.

It is straightforward to represent $\otimes_a$ using the modal operators $\Box$ and K_a as follows:

$$\otimes_a(\pi_x(\mathcal{O}), \pi_x(P(\mathrm{a}))) := \Box(\pi_x(\mathcal{O}) \rightarrow \pi_x(P(\mathrm{a}))) \wedge \neg K_a \neg(\pi_x(\mathcal{O}) \rightarrow \neg\pi_x(P(\mathrm{a})).$$

We now present an example illustrating a possible application of the general scheme for bias given in Def. 5. Consider the model $\mathcal{M}^{bias}$ (see Fig. 1). In this model, agents a and b are subject to bias with respect to $P(\mathrm{a})$, whereas agent c is not. For every world w in $\mathcal{M}^{bias}$, we have $\mathcal{M}, w \models_\sigma \pi_x(\mathcal{O}) \rightarrow \pi_x(P(\mathrm{a}))$, which captures the normative constraint.[2] However, while agent c does not consider it epistemically possible that $\neg\pi_x(P(\mathrm{a}))$ together with the ontology $\mathcal{O}$ hold, agents a and b do so. The underlying reasons, however, may differ. In particular, agent a considers w_2 possible. In that world, the ontology differs from the one specified by the normative constraint, which leads to the outcome $\neg\pi_x(P(\mathrm{a}))$. When analyzed in more detail, this situation can be associated with sampling bias, which arises when the source of an inference lacks observations about a relevant class of objects. Agent b, by contrast, does not consider w_2 possible; hence, in all worlds accessible to b, $\pi_x(P(\mathrm{a}))$ holds. Nevertheless, in w_3, the output $\pi_x(P(\mathrm{a}))$ is not obtained under $\mathcal{O}$. This suggests that agent b may be subject to a different

[2]Notice that the norm takes the form of an implication, where the condition of ideal functioning, $\pi_x(\mathcal{O})$, appears as the antecedent. Consequently, it is not necessary that an agent conceives the ontology as holding in every epistemically possible situation.

form of bias, namely deployment bias, which occurs when an ML system is applied in contexts that differ significantly from those in which it was trained.

By integrating normative and epistemic considerations, the framework proposed here provides a novel perspective on bias as a phenomenon that affects not only the outputs of ML systems but also the attitudes of the agents who rely on them. This perspective may inform future approaches to bias detection and mitigation, ultimately contributing to the development of more trustworthy AI systems. Future work will refine the general framework by formally characterizing specific forms of bias and investigating their manifestation in concrete ML applications.

Acknowledgments

The authors would like to thank Roberta Ferrario and Daniele Porello for their valuable comments on an earlier version of this work. They are also grateful to the anonymous reviewers of DL 2026 for their constructive feedback and helpful suggestions. Mattia Petrolo acknowledges the financial support of FCT – Fundação para a Ciência e a Tecnologia through project "Algorithm and Knowledge - Toward an Epistemic Approach to Artifcial Intelligence", ref. 2022.08338.CEECIND/CP1722/CT0013, DOI 10.54499/2022.08338.CEECIND/CP1722/CT0013 (https://doi.org/10.54499/2022.08338.CEECIND/CP1722/CT0013), and the LASIGE Research Unit, ref. UID/408/2025. Ekaterina Kubyshkina's research is supported by the University of Milan through the project "Foundations of Fair and Trustworthy AI" and by the Department of Philosophy "Piero Martinetti" of the University of Milan under the project "Departments of Excellence 2023-2027" awarded by the Ministry of University and Research (MUR).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT based on GPT-5.5 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. C. Khadir, H. Aliane, A. Guessoum, Ontology learning: Grand tour and challenges, Computer Science Review 39 (2021).

[2] D. Dou, H. Wang, H. Liu, Semantic data mining: A survey of ontology-based approaches, in: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, 2015, pp. 244–251.

[3] P. Hitzler, M. K. Sarker (Eds.), Neuro-Symbolic Artificial Intelligence: The State of the Art, IOS Press, 2022.

[4] R. Ferrario, D. Porello, E. Bottazzi, C. De Florio, M. Fumagalli, Taming the sea of errors: An ontological study of biases in dolce, in: Formal Ontology in Information Systems FOIS 2025, IOS Press, Netherlands, 2026, pp. 64–78. doi:10.3233/faia250484.

[5] T. Kelly, Bias, norms, introspection, and the bias blind spot, Philosophy and Phenomenological Research 108 (2024) 81–105.

[6] F. Baader, I. Horrocks, C. Lutz, U. Sattler, An introduction to Description Logic, Cambridge University Press, 2017.

[7] F. M. Donini, M. Lenzerini, D. Nardi, W. Nutt, A. Schaerf, An epistemic operator for description logics, Artificial Intelligence (1998) 225–274.

[8] A. Mehdi, S. Rudolph, Revisiting semantics for epistemic extensions of description logics, in: AAAI'11: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011, pp. 241–246.

[9] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, EQL-Lite: Effective first-order query processing in description logics, 2007, pp. 274–279.