Formal Ontology in Information Systems
T. Prince Sales et al. (Eds.)

doi:10.3233/FAIA250484

Taming the Sea of Errors: An Ontological Study of Biases in DOLCE

Roberta FERRARIO [a,1], Daniele PORELLO [b], Emanuele BOTTAZZI [a], Ciro DE FLORIO [c] and Mattia FUMAGALLI [d]

[a] *Institute of Cognitive Sciences and Technologies – CNR, Trento, Italy*
[b] *University of Genoa, Genoa, Italy*
[c] *Università Cattolica, Milan, Italy*
[d] *Free University of Bozen-Bolzano, Bolzano, Italy*

Abstract. In this paper, we present a preliminary ontology of bias based on the DOLCE foundational ontology. The main reason for devising such an endeavour is to make explicit the ontological assumptions behind the use of terms indicating the elements composing a biased outcome. Firstly, we discuss what the object of a bias is —namely, the entity that might be deemed biased, which we identify with situated inferences, i.e. propositional contents that can be asserted by some (human or artificial) agent from other propositional contents. We will thus categorise in DOLCE various types of biases as concepts that classify situated inferences. The content of such inferences is then associated with the following elements: *i*) the agent responsible for drawing the conclusion, *ii*) the objects and *iii*) the concepts used in the premises and in the conclusion of the inference, *iv*) the time when the inference takes place. These ingredients will serve to trace the origin of what we shall call a biased inference back to any of the above elements, relating some of the biases present in the literature to these ontologically founded elements.

Keywords. representation of bias, observation, inference, Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE).

1. Introduction

The awareness that our judgments can be systematically flawed—whether due to cultural blind spots, ingrained prejudices, or the limitations of our tools (conceptual, measuring, or computational)—is by no means new. Indeed, from the earliest days of human thought, we have carried this inherent shortcoming, accompanied by an ever-growing body of examples illustrating our awareness of these limits. In today's highly technological and digitally dependent society, it should come as no surprise that we devote considerable effort to examining and mitigating what we now call 'bias'. The real challenge, however, lies in bringing some order to the vast array of circumstances in which such biases occur.

The word 'bias' appears in discourses at all levels, from casual conversations to scientific research across fields. This widespread use reflects how bias emerges in various contexts—everyday interactions, legal judgments, cognitive sciences, and machine

[1]Corresponding Author: Roberta Ferrario. E-mail: roberta.ferrario@cnr.it

learning (ML) and its applications. In everyday conversations, accusations of bias typically refer to feelings, opinions, or decisions sustained by unfair judgments or flawed reasoning. On the other hand, admissions are usually meant to express a preference or an interest, not necessarily ascribing a positive or negative value.

In legal contexts, judicial bias occurs when someone is so committed to a specific outcome that they disregard contrary evidence. Cognitive sciences and psychology identify biases as characteristic patterns in information processing that lead to inaccurate outcomes [1,2]. Qualitative sociological research has traditionally examined prejudices that cause inaccurate interpretations of social phenomena [3]. Quantitative sociological studies and statistics-based sciences view biases as systematic deviations from valid measurements that make statistics unrepresentative of that population [4]. In computer science, particularly in ML, bias typically refers to systematic errors in model predictions relative to a reference considered a correct output, that is, the 'ground truth'. Since ML uses statistics to identify patterns in the data, most of the kinds of bias studied in ML are also inherited from statistics.

Given these diverse manifestations, we propose that a philosophical, epistemologically-focused approach provides a fruitful starting point for understanding bias. We frame this approach in terms of applied ontology methods, seeking a level of abstraction capable of capturing how different disciplines characterise systematic deviations. This approach allows us to begin this work of reordering starting from one of the aspects most closely related to ontology: classification tasks.[2] The crux of the matter lies in understanding how knowledge acquisition occurs: an agent—whether human or artificial—accumulates knowledge based on certain observations, and this knowledge, regardless of its veracity, enables the agent to classify entities in one way rather than another. Our study examines how distortions can emerge at various points within this complex interaction between observations, acquired knowledge, and resulting classifications. In a nutshell, we will classify different types of bias according to these interrelations. As human and artificial outputs increasingly converge, the need to identify systematically skewed judgments becomes more pressing. Therefore, we focus on the domains of ML and cognitive science, where biases share socio-technical dimensions—whether directly embedded by humans or emerging from seemingly neutral computational processes.

The paper is structured as follows. Section 2 discusses the contribution of ontologies to bias analysis. Section 3 explores different aspects of bias, first examining philosophical perspectives (3.1), and ML formulations (3.2). Section 4 presents a conceptual analysis of bias based on some philosophical accounts (4.1) and a taxonomy synthesizing various classifications of biases in ML ordered according to conceptual criteria (4.2). Section 5 introduces our proposal, aimed at providing an axiomatisation of the taxonomy; it begins with an overview of the DOLCE ontology (5.1), then it explains how biases are categorised within it (5.2), develops the concepts of observation and inference (5.3) to define biases as concepts that classify inferences (5.4). Finally, Section 6 sketches some possible future directions of the work.

[2] Bias is also discussed in relation to tasks that are not strictly classificatory, such as in Generative AI [5].

2. The Contribution of Ontologies

Before engaging in conceptual analysis and modeling, it is worth motivating the use of formal ontologies to complement other techniques in handling biases in ML.

To mitigate biases emerging in the use of ML systems, they need to be first uncovered, identified and represented. Given the diverse kinds of bias that can emerge at different stages of the ML systems' pipeline [6], the representation task is paramount, though often overlooked. Formal ontologies are thus a powerful tool for understanding and documenting biases, as well as providing a shared vocabulary. This is also a necessary step to enable the mitigation of biases.

Although representation is surely the main purpose of ontologies, they might also be used in conjunction with other techniques to interpret the training data and to reveal inconsistencies in both automatic and manual annotations. They can be used to provide a semantic interpretation of the components of datasets and to improve the expressivity of the queries directed to such datasets [7].

Unfortunately, not many ontologies of biases are available nowadays. A notable exception is the *Fairness Metrics Ontology* (FMO) [8], which has the purpose of evaluating ML models with respect to fairness. In such an approach, fairness is defined as non-discrimination and, since discrimination may be the consequence of a bias, the ontology associates to some biases identified in the literature a corresponding fairness metrics, which measures the degree of success in contrasting the bias. So, by querying the ontology, one may receive a recommendation on which bias to address, given some specific 'fairness scores'.

A more recent work may be found in [6], which presents the *Doc-BiasO* ontology that, by reusing other ontologies, identifies biases, assigns a quantitative value to them and associates different risks to different biases in different contexts.

Although these ontologies are very useful tools to direct the application of mitigation measures, they take the notion of bias as given, without enquiring what biases are, which are the elements of the ML system with which they are related and how.

We believe that a conceptual characterisation of an umbrella concept of bias could provide the basis for implementing ML strategies to mitigate biases.

3. Aspects of Biases

3.1. The Philosophical Debate

Philosophical studies on bias are mainly directed towards human (cognitive or social) bias [9], while mostly neglecting the analysis of bias resulting as output of a ML system (what in this debate has been often called 'algorithmic bias'), with few exceptions (see, for instance, [10]).

However, it is extremely interesting to look into this literature as some of the distinctions that can be found there can shed light on the ML domain as well. Furthermore, given the socio-technical character of bias, it seems desirable to single out a notion encompassing such diverse uses of the term.

According to a quite widespread philosophical characterisation, two major families of theories on biases are present: the *functional* and *norm-theoretic* views. According to the former, “Bias arises to overcome uncertainty, taking underdetermining evidential states as input and producing a determinate “best guess about reality”. [11]; i.e., they are used to choose a preferred option over a set of equally fitting alternatives [12].

An interesting functional account is that proposed by Gabbrielle Johnson [13], who sees them as the combination of a (typically stereotyped) bias-input and a bias-construct, an operation on this content (for example, a universal quantification or a statistical generalisation), which together produce a bias-output. Bias-input, bias-construct and bias-output are in this approach committal *mental states* with a *propositional content*. Thus, bias is not a mental state in itself, rather, it is a complex *construct of mental states*.

Norm-theoretic views, instead, see biases as *systematic deviations from a norm*. One of the proponents of such a view is Thomas Kelly [14], who also offers different interpretations of what a ‘genuine’ norm is without committing to any of these interpretations. Depending on who attributes bias, the norm referred to could be truth, accuracy, value maximisation, evidence, a moral norm, or justice. The term ‘bias’ is frequently understood negatively, often associated with stereotypes, prejudice, and discrimination. Proponents of the norm-theoretic view—where bias is seen as a deviation from a norm—tend to assume implicitly that the norm is positive, thereby casting the deviation (bias) as negative. But the term ‘bias’ itself is not inherently negative. Deviating systematically from an unjust law could be viewed as a *positive* bias. Of course, one could argue that in the case of an unjust law, it is the *law* that is biased—e.g., against a higher ethical norm. However, in order to model bias ontologically, we do not believe it is necessary to take a stance on which norm is superior to another. In fact, it might be interesting to model the widespread practice of correcting one bias by introducing another—as in the case of gender quotas in hiring and representation. Whether ethically justified or not, such corrective practices are prevalent in our society and warrant ontological analysis.

Although the two approaches are often regarded as antithetical, we believe that they actually pertain to two conceptual dimensions of bias and implicitly refer to two distinct dimensions of normativity. However, an exploration of this point lies beyond the scope of the present work.

A further distinction analysed in philosophy is that between explicit and implicit bias. While the former are seen as conscious stereotyped beliefs that are available for introspection, the latter are not accessible by conscience and are often automatic. Since there is a widespread consensus on the nature and role of explicit biases, the debate concentrates rather on implicit ones and on whether they have representational content.[3]

3.2. Aspects of Biases in Machine Learning

Compared to philosophical approaches, computer science literature focuses on identifying bias sources with the more pragmatic aim of developing mitigation strategies, although philosophical distinctions remain discernible, if implicit.

Several definitions from literature illustrate key perspectives: “bias refers to a deviation from a standard” [16]; “algorithmic bias is simply systematic deviation in algorithm output, performance, or impact, relative to some norm or standard” [17]; algorithmic bias occurs when “the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason” [18]; or “the systematic tendency in a model to favor one demographic group/individual over another” [19].

[3]For a comprehensive account, see [9] and [15].

We may observe that the two former definitions align with norm-theoretic views referencing statistical standards, while the latter emphasize negative social consequences and implicitly reference ethical norms.

Moreover, these definitions may misleadingly suggest that all ML biases are detrimental. Some biases are instead deliberately embedded to ensure positive outcomes, as [16] notes, with 'ethical governors' that modify outputs to prioritise ethical choices. Similarly, [20] observes that "bias is necessary to setup a machine learning task", since learning from all observable reality would be impossible—echoing the uncertainty reduction function of bias in philosophy and even more the practice of correcting one bias by introducing another. That is, the implicit/explicit distinction also appears in ML contexts: some biases are deliberately introduced, such as excluding protected variables for ethical reasons [16], while others emerge unintentionally when proxy variables correlate with protected attributes, as in the notorious COMPAS case [21].

4. Conceptual Analysis and Taxonomy of Bias

4.1. Leveraging Metaphysical Theories to Study Biases

The philosophical debate on biases includes a branch specifically concerned with their metaphysical characterisation. While most studies focus on implicit biases, their framework can extend to explicit ones as well. Across various positions, biases are generally viewed as attributes, though perspectives differ regarding the metaphysical entities they qualify.

The predominant view considers biases as attributes of mental attitudes—either simple beliefs [22] or complex cognitive schemas [13]. Some scholars treat all mental attitudes as propositional [22,23], while others distinguish between propositional beliefs and aliefs [24], which combine representational, affective, and behavioural aspects. Alternative approaches [25] interpret biases as attributes of mental processes within a dual-system cognitive framework [26], distinguishing impulsive associative processes and reflective propositional ones.

For our purposes, we interpret bias in relation to entities with propositional structure, but we focus on publicly observable aspects rather than internal processes. This approach allows to address both human and ML biases, despite the black-box nature of many ML systems. We evaluate bias in what is publicly available—behaviours, judgments, and outputs—rather than in thoughts or internal computations.

'Judgment' encompasses both the mental act of judging and its expression [27]. Asserting a claim involves formulating propositional content but also performing a linguistic act[4] [28]. Such assertions express an agent's perspective on observations, functioning both as psychological phenomena (or computational elaborations) and communicative deeds. Bias can occur in any of these dimensions—systematic skews relative to available data or data that could have been gathered. Recognising such deviations requires comparison against *standards* or *desiderata* through evaluative acts that are inherently *conceptual*. When we classify an entity under a particular concept, we appeal to reasons grounded in specific properties that the entity exhibits. Bias, then, is for us a kind

[4] Our approach is inspired by speech acts theories, but in the current paper we focus solely on acts of claiming.

of *concept* that classifies the process leading to *infer* certain acts of classification by an agent.

4.2. A Taxonomy of ML-Inspired Biases

Drawing inspiration from existing ML bias taxonomies, we develop our own classification based on ontological distinctions rather than pipeline stages. While most literature organises biases by their emergence point in ML systems [19,20,21], we concentrate on the ontological elements involved.

A key distinction highlighted by [16] is between statistical bias (deviation from statistical standards) and moral bias (deviation from ethical norms). Crucially, "the very same thing can be biased according to one standard, but not according to another." This emphasises our central point: bias evaluation always requires reference to some standard. We therefore distinguish between statistical and social[5] bias based on their reference standard—statistical datasets or social/ethical/juridical norms, both conceived of as complex observations.

Biases in our framework are concepts classifying *inferences*, involving the agent producing (expressing) the observation, the observation itself, the set of previous observations that ground it and the time of production. We will thus attribute a bias to an observation on the basis of the choice of one of the elements composing the production which we deem incorrect (with respect to a certain statistical or social standard). Our framework will thus distinguish:

Sampling bias[6] that occurs when the choice of the observations in the source dataset is incorrect. This may be statistical (e.g. unrepresentative samples) or social (e.g. deliberately altering sample distribution to address discrimination).

Specification bias[7] relates to incorrect conceptual specification in the source dataset or in the discrete classes used for classification. This includes inappropriate feature selection and class labeling, manifesting itself statistically (omitting relevant features) or socially (deliberately excluding sensitive attributes).

Inductive bias[8] concerns the agent's inductive process. For ML systems, this encompasses architectural choices (hidden layers, activation functions, algorithms, loss functions), while maintaining an abstract black-box perspective.[9] Examples include statistical regularisation parameters or socially motivated alterations to system behaviour in sensitive contexts, like in the case of a self-driving system whose behaviour is purportedly modified when the vehicle is approaching an area with schools.

Deployment bias[10], which arises when a ML system is deployed in a scenario that differs from that of training, as in the case of a self-driving system trained in the US, where vehicles travel on the right side of the road and deployed in the UK, where they travel on the left.

[5] This term is more inclusive than 'moral'.

[6] Term used in [20,21], while [16] calls it "training bias" and [19] "bias in representativeness of the sample".

[7] See [20]. It aligns with the "algorithmic focus bias" in [16] and covers cases under the "data to algorithm bias" category in [21].

[8] This term is taken from [20], though used differently: while they refer to a limited search space for functions and use "hyper-link parameters bias" for what we mean here, we found 'inductive' more explanatory and suitable also in the human case. It likely corresponds to what [16] calls "algorithmic processing bias" and what [21] refers to simply as "algorithmic bias".

[9] At this level of analysis, it is useful to abstract away also from human neurophysiology and psychology.

[10] This is the only term we did not inherit from [20], but decided to adopt the term used by [19].

5. Sketching a Proposal

5.1. *A Brief Introduction to* DOLCE

DOLCE is a foundational ontology that adopts a descriptive rather than prescriptive metaphysics, aiming to capture the commonsense conceptualisations reflected in natural language, human cognition, and social practices [29,30]. The core ontology is designed to be extended with domain-specific modules and has been applied in a wide range of domains—as varied as healthcare and robotics. DOLCE organises entities into several categories that are subsumed under four fundamental ones:
Endurants (Objects): These are entities that persist through time while maintaining their identity despite changes in properties or parts. Endurants are wholly present at every moment of their existence.
Perdurants (Events): In contrast, perdurants are entities that unfold over time and are characterised by their temporal parts.
Qualities: These are properties or attributes that can be perceived or measured, inhering specifically to an individual object.
Abstract Entities: These are entities that lack spatial or temporal qualities, such as mathematical objects.

Furthermore, [31] introduces the notion of **(Social) Concept**, a type of entity used to dynamically classify other entities. In this paper, we leverage the notion of a concept to identify the properties that an entity must possess in order to be classified as biased.

5.2. *Categorising Biases in* DOLCE

Firstly, we discuss what the object of a bias is—namely, the entity that might be deemed biased in DOLCE. We view biases as applying to inferences, understood here as propositional contents that are asserted by some agent —human, social or artificial—with respect to other propositional contents. We do not commit to any view of correctness of inferences, for instance, concluding that the next raven the agent *i* sees is white on the ground that *i* has already seen two white ravens is an inference as well[11].

To include biases within the taxonomy of DOLCE, we shall first include inferences as particulars of DOLCE. For this purpose, we use the notion of an *observation* introduced in [32]. Thus, an inference will be rendered as a way to obtain an observation using other observations. We will categorise various types of biases as various ways of classifying inferences, so in DOLCE, the types of biases are included as *concepts* that classify inferences. For instance, agent *i*'s inference about ravens might be deemed biased, thus classified under a certain concept (type) of bias. Figure 1 illustrates an excerpt of the taxonomy of DOLCE with biases, observations, and inferences.

The theory of observations, inferences, and biases developed here extends the first-order logic version of DOLCE.[12]

[11]Notice that we may have statistically correct inferences that we deem biased on ethical grounds. In the COMPAS case [21], for instance, the claim that individuals who live in a family context where other members have committed crimes are more likely to reoffend is probably statistically accurate, but ethically biased. Conversely, we may have statistically unbiased inferences that are still considered ethically biased (e.g., affirmative action or positive discrimination).

[12]For the version of DOLCE, see `https://github.com/appliedontolab/DOLCE`. Our

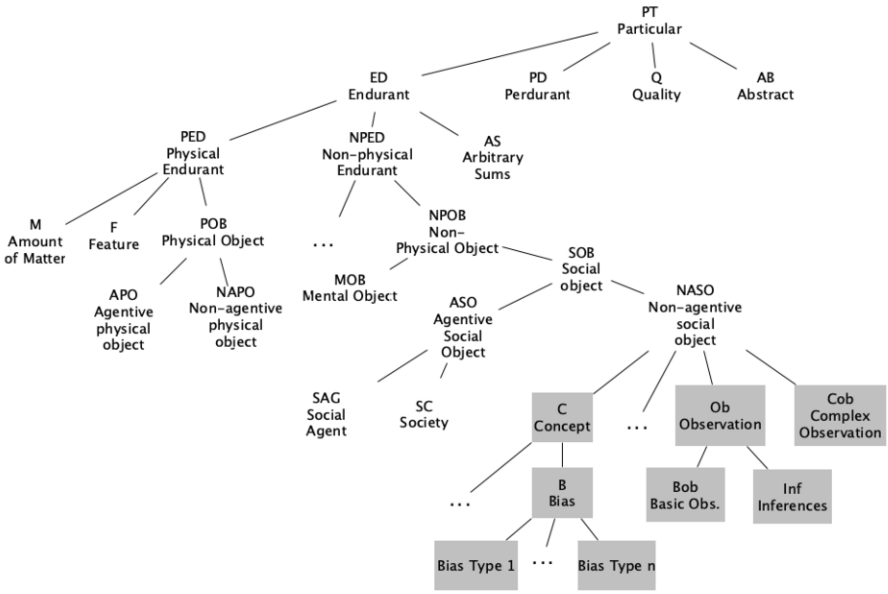

Figure 1. An excerpt of the taxonomy of DOLCE with observations, inferences, and biases

5.3. *Observations and Inferences*

Observations are understood as registrations of publicly expressed propositional contents. For the sake of simplification, we restrict our current analysis to classification contents, i.e. contents of the form 'the entity x is classified under the concept c at time t'. In DOLCE, we can use the *classification* relation $\mathsf{cf}(x,c,t)$ to represent such contents, see [30] and [31]. Observations are not *factual*, that is, the sole presence or existence of an observation does not entail that its content (the classification proposition) is true; namely we are interested in wrong observations as well.

We now turn to discussing how to categorise observations within the taxonomy of DOLCE. As situated proposals of classifications, observations are not categorised under Abstract in DOLCE, as abstracts are intended to be out of space and time. In this respect, abstracts include propositions or facts [29], not observations that occur in time. On the other hand, if we classify observations under Perdurants (e.g. events), we cannot properly refer to two occurrences of the same observation since they would instead be two distinct observations, i.e., two events happening in different moments. We propose to classify observations (Ob) under NonPhysicalEndurant (Nped) in DOLCE. This amounts to assuming that observations are in time, but we abstract from their spatial location.

extension and the documentation is available at https://github.com/appliedontolab/DOLCE-Bias through the permanent w3id link https://github.com/perma-id/w3id.org/tree/0677aafbdd17463e27e229cd96a4f839c88b3ea1/dolce-bias. We designed our theory for running model finders to test the consistency of our extension. In particular, we used Mace4 (see https://www.cs.unm.edu/~mccune/mace4/manual/2009-11A/) to obtain a model of the theory. We restricted to the sub-theory of DOLCE containing axioms about endurants to enable Mace4 to effectively produce a populated model of the theory.

Moreover, the same observation may occur at different times in various contexts of other observations. More specifically, we classify them as NonAgentiveSocialObjects (Naso), to highlight the public, inter-subjective nature of observations, cf. axiom (a1). Moreover, observations are distinguished from other Naso, specifically from concepts, (a3), which are entities that classify other entities, whereas observations do not.

We believe this choice aligns with everyday language: an observation can be repeated, compared, or even refuted. If the process of observing can be considered a perdurant—an activity unfolding over time, as when we say "the observation of the stellar event lasted for three hours"—then the product of that act of observing behaves more like an endurant, that is, an entity that exists wholly at each moment it is present.

We discuss an excerpt of the theory, while we refer to our repository for the exhaustive list of axioms.

a1 $\mathsf{Ob}(x) \rightarrow \mathsf{Naso}(x)$

a2 $\mathsf{Cob}(x) \rightarrow \mathsf{Naso}(x)$

a3 $\neg\exists x(\mathsf{Ob}(x) \wedge \mathsf{C}(x))$

a4 $\neg\exists x(\mathsf{Cob}(x) \wedge \mathsf{Ob}(x))$

a5 $\neg\exists x(\mathsf{Cob}(x) \wedge \mathsf{C}(x))$

a6 $\neg\exists x(\mathsf{Bob}(x) \wedge \mathsf{Inf}(x))$

a7 $\mathsf{Bob}(x) \leftrightarrow \mathsf{Ob}(x) \wedge \forall t(\mathsf{T}(t) \rightarrow \neg\exists y(\mathsf{Ob}(y) \wedge \mathsf{tpp}(y,x,t)))$

a8 $\mathsf{Cob}(x) \rightarrow \exists yzt(\mathsf{Bob}(y) \wedge (\mathsf{Bob}(z) \vee \mathsf{Cob}(z)) \wedge\ y \neq z\ \wedge \mathsf{sumt}(x,y,z,t))$

a9 $\mathsf{Bob}(x) \leftrightarrow \exists! yzct\ (\mathsf{hasAgent}(x,y) \wedge \mathsf{hasObject}(x,z) \wedge \mathsf{hasConcept}(x,c) \wedge \mathsf{hasTime}(x,t))$

a10 $\mathsf{hasAgent}(x,y) \rightarrow \mathsf{Ob}(x) \wedge (\mathsf{Apo}(y) \vee \mathsf{Aso}(y))$

a11 $\mathsf{hasObject}(x,y) \rightarrow \mathsf{Bob}(x) \wedge \mathsf{Ed}(y)$.

a12 $\mathsf{hasConcept}(x,y) \rightarrow \mathsf{Bob}(x) \wedge \mathsf{C}(y)$.

a13 $\mathsf{hasTime}(x,y) \rightarrow \mathsf{Bob}(x) \wedge \mathsf{T}(y)$

a14 $(\bigwedge_{i\in\{1,2\}}(\mathsf{hasAgent}(x_i,y) \wedge \mathsf{hasObject}(x_i,z) \wedge \mathsf{hasConcept}(x_i,c) \wedge \mathsf{hasTime}(x_i,t))) \rightarrow x_1 = x_2$

a15 $\mathsf{Inf}(x) \leftrightarrow \mathsf{Ob}(x) \wedge \exists! yzv(\mathsf{hasAgent}(x,y) \wedge \mathsf{hasSource}(x,z) \wedge \mathsf{hasTarget}(x,v))$

a16 $\mathsf{hasSource}(x,y) \rightarrow \mathsf{Inf}(x) \wedge (\mathsf{Bob}(y) \vee \mathsf{Cob}(y))$

a17 $\mathsf{hasTarget}(x,y) \rightarrow \mathsf{Inf}(x) \wedge \mathsf{Bob}(y)$

a18 $\mathsf{Inf}(x) \wedge \mathsf{hasSource}(x,y) \wedge \mathsf{hasTarget}(x,z) \wedge \mathsf{pre}(x,t) \rightarrow \mathsf{pre}(y,t) \wedge \mathsf{pre}(z,t)$

The class of observations is divided into: *i*) *basic* observations (Bob), which represent simple statements about classifications and *ii*) *inferences* (Inf), that represent the assertion of basic observations on the ground of other basic observations.

Basic observations do not have other observations as temporary proper parts (tpp in DOLCE), cf. (a7), while we introduce the class of *complex* observations (Cob), which coincide with mereological temporary sums of distinct basic observations (cf. (a8), where sumt indicates the temporary sum of DOLCE). Notice that the class Cob is not included under Ob, as we shall associate a single agent, who is responsible for the observation, to Bob and Inf but not to Cob, which may be provided by several agents.[13]

A basic observation represents a statement of the form "*i* states that Tom is a cat at time *t*". A complex observation represents a collection of statement, e.g.,"*i* states that Tom has whiskers at *t*" and"*j* states that Tom meows at *t*". A complex observation can represent a *dataset* of classifications. An inference represents a conditional assertion

[13]The term 'complex observations' is motivated to remind the label used in [32], while inferences are termed 'data productions' there. Notice that the class Cob is not strictly required: DOLCE enables mereological sums of endurants, which are themselves endurants. The modification with respect to the original DOLCE is that the sum of observations, which are Naso, is still categorised as a Naso here.

"given that i states that Tom has whiskers at t and j states that Tom meows at t, h infers that Tom is a cat at t".

Basic observations Bob are associated with the following entities, which provide the content of an observation, (a9): *i*) the agent responsible for the observation, *ii*) the object classified in the observation, *iii*) the concept used in the observation, *iv*) the time when the classification is asserted to hold.[14] An observation is associated to its ingredients by means of the primitive relations in (a11), (a12), and (a13). Moreover, basic observations coincide when they have the same agent, object, concept, and time, cf. (a14).

The object classified in an observation is here any endurant (Ed in (a11)). The agent who is responsible for the observation is an AgentivePhysicalObject (including human agents as well as artificial agents, such as classification algorithms, in a Dennettian way) or an AgentiveSocialObject (a jury, a company), cf. (a10). The time of the classification is a time interval of DOLCE (T in (a13)).

Inferences Inf are associated with the following entities: *i*) a single agent, who is responsible for the inference, *ii*) a unique *source*, which is a mereological sum of one or more basic observations, and *iii*) a unique *target*, which is a basic observation. In our intended interpretation, the source observations are actually used to produce the target observation. Accordingly, we assume that, in an inference, both source and target are present when the inference is present, (a18). Naturally, this target observation—produced from the source observations—is not necessarily derived through a correct inferential process; indeed, the relation between these relata is precisely the object of our bias analysis.[15]

Due to the existence and unicity assumptions enforced by axiom (a9) and (a15), we may introduce the following functional terms to simplify the subsequent formulas. We indicate by $\mathsf{ob}(x)$, $\mathsf{cn}(x)$, and $\mathsf{tm}(x)$, the object, the concept and the time (respectively) of a Bob. We indicate by $\mathsf{ag}(x)$ the agent of a Bob or an Inf. By $\mathsf{sr}(x)$ and $\mathsf{tr}(x)$, we indicate the source and the target (respectively) of an Inf. These ingredients will serve to trace the origin of a biased inference back to any of the above elements.

5.4. *Biases as Concepts in* DOLCE

According to our proposal, biases classify inferences. To categorise biases, we use the class of *concepts* (C) of DOLCE. Therefore, a bias is a type of concept: $\mathsf{B}(x)$ indicates that x is a bias. We write formulas of the form: an inference p is classified as a certain bias b at a certain time t: $\mathsf{cf}(p,b,t)$, cf. (a19). Accordingly, DOLCE infers $\mathsf{B}(x) \to \mathsf{C}(x)$.

a19 $\mathsf{B}(x) \to \exists pt(\mathsf{Inf}(p) \land \mathsf{T}(t) \land \mathsf{cf}(p,x,t))$

a20 $\mathsf{requires}(x,y) \to \mathsf{C}(x) \land \mathsf{C}(y)$

a21 $\mathsf{requires}(c,c_r)$

f1 $\mathsf{requires}(y,y') \leftrightarrow \forall xt(\mathsf{cf}(x,y,t) \leftrightarrow \mathsf{cf}(x,y',t))$

[14]Observations and inferences are endurants, so they are present in time in DOLCE. The model allows for observations and inferences, present at time t, to concern classifications referring to the past or future relative to t. For example, at t the predictive observation about a future time t' "the weather is sunny at t'" may be present and may still be present at a subsequent time t''. We do not enter here into the discussion about the persistence conditions of observations; we simply note that, as recorded assertions, observations might persist even when the agent is not present — for example, when one uses another's observations to infer an observation.

[15]Here, we do not assume that inferences with same sources, targets and agents must coincide, as there might be more ways of inferring the target from the source.

Several notions of biases use the standard of an 'ideal inference'. To render, abstractly, this idea, we introduce a concept c_r, which is required to produce a correct classification under concept c. The relationship between c and c_r is represented by the primitive relation requires between concepts. We do not approach here the analysis of this relation; as we discussed, it is quite heterogeneous and in principle we would need many requirement relations: it might be construed in terms of sufficient or necessary conditions for classification, or in terms of a strong statistical correlation, or causality, or even as a normative legal requirement.[16] Also, as a simplification move, we assume here that requirements are stable through time while, in principle, they might vary (e.g., when new knowledge becomes available, or when the jurisdiction changes).

Formula (f1) provides an example of a requirement in terms of (constant in time) necessary and sufficient conditions. We could also admit several types of dependencies, e.g. statistical, normative, factual, etc., that would not meet formula (f1).

Axiom (a21) exemplifies a list of assertion axioms, one for each concept c that we are discussing. So, here c and c_r are individual constants and the requirement statements are taken as given.

In the following paragraphs, we propose definitions for several biases. We will use (decorated) individual constants, $b_{Ty,h_1,\dots,h_n}$, to indicate an instance of a particular type of bias Ty, where the parameters h_i illustrate the elements that may be required to define a bias (e.g., a bias might be so only with respect to certain concepts). A class $B_{Ty}(x)$ of biases can then be defined by grouping together all those instances that satisfy a given condition. To define a bias, we propose extensional definitions, providing necessary and sufficient conditions for classifying inferences as biased.

Finally, to simplify the notation, when c is a concept, we write $\bar{c}$ to indicate a concept that is incompatible with c, i.e., such that $\mathsf{cf}(x,\bar{c},t) \rightarrow \neg\mathsf{cf}(x,c,t)$.

Example 4.1. As an example of the functioning of our framework, we approach a well-known type of bias introduced by Daniel Kanheman in Cognitive Science, dubbed the 'framing effect'. Kahneman notes on several occasions [1,33] how in some situations it is quite clear that our judgments (in our case, classifications) diverge from a true value or, at least, from a probabilistically adequate one.

Consider the following type of bias, similar to those discussed by Kahneman.[17] Suppose that an agent i classifies *healthy* foods (c in (a22)) on the basis of their *green wrap* (c'). Even when agent i is presented with evidence that might produce a correct classification (here a concept c_r which is required for c, e.g. *low calories food*), agent i's inference still classifies that green-wrapped food as healthy. Moreover, i classifies food as non-healthy whenever the information about green wraps is absent from its source observations (e.g. when *red wrap* is present in the source). We note that this type of bias concerns two concepts c (healthy) and c' (green-wrapped) and an agent i, i.e., i could be not biased w.r.t. other concepts (here again c, c' and i are individual constants).

We denote the an instance of such bias by means of the individual constant $b_{K,c,c',i}$, to indicate the dependence on c', c, and i.

[16] In principle, we might have several conditions c_{r_i} for correctly classifying under c. As a simplification, we do not explicitly model that case. Notice also that a comparable notion of requirement is used in [31] to express conditions on classifications.

[17] The classical paper for the framing effect is [34]. For an application to marketing and to consumer behaviour, see [35]. The latter work takes into account the framing effect concerning the evaluation of meat; we used, instead, a toy example, similar in spirit.

a22 $\mathsf{cf}(p, b_{K,c,c',i}, t) \leftrightarrow$
$\mathsf{pre}(p,t) \wedge \mathsf{ag}(p) = i \wedge \exists o'(\mathsf{tp}(o', \mathsf{sr}(p), t) \wedge \mathsf{cn}(o') = c') \wedge \mathsf{cn}(\mathsf{tr}(p)) = c$
$\wedge$
$\forall p' o' o''(\mathsf{ag}(p') = i \wedge \mathsf{tp}(o', \mathsf{sr}(p'), t) \wedge \mathsf{tp}(o'', \mathsf{sr}(p'), t) \wedge \mathsf{cn}(o') = c' \wedge \mathsf{cn}(o'') = \bar{c}_r \wedge$
$\mathsf{ob}(o') = \mathsf{ob}(o'') \wedge \mathsf{cn}(\mathsf{tr}(p')) = c \rightarrow \mathsf{ob}(\mathsf{tr}(p')) = \mathsf{ob}(o'))$
$\wedge$
$\forall p' o'(\mathsf{ag}(p') = i \wedge \mathsf{tp}(o', \mathsf{sr}(p'), t) \wedge \mathsf{cn}(o') = \bar{c}' \wedge \mathsf{cn}(\mathsf{tr}(p')) = \bar{c} \rightarrow \mathsf{ob}(\mathsf{tr}(p')) = \mathsf{ob}(o'))$

Definition (a22) states that an inference p is classified under the bias $b_{k,c,c',i}$ iff *i*) p is an inference of agent i, p uses c' in its source, and p uses c in its target, *ii*) for every inference p' of agent i whose source classifies an object under both c' and not c_r (i.e., $\bar{c}_r$), the target of p' still classifies the object under c, and *iii*) for every inference p' of agent i whose source classifies an object under not c' ($\bar{c}'$), the target of p' classifies the object under not c (i.e., $\bar{c}$).[18].

In the following paragraphs, we offer general definitions of the types of bias discussed in Section 4.2.

5.4.1. Sampling Bias

A sampling bias ($b_{Sa,c,i}$) affects the source of an inference for a certain target concept (c) and agent (i). I.e., a dataset can be biased for the target classifications of x under c, while not biased for other classifications. We indicate by $B_{Sa,c,i}$ the class of these biases. The idea is that an inference is biased in this sense if its source lacks observations about an important class of objects (indicated by means of the concept cl).

a23 $\mathsf{cf}(p, b_{Sa,c,i}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge$
$\mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \forall o(\mathsf{cf}(\mathsf{obj}(o), cl, t) \rightarrow \neg \mathsf{tp}(o, \mathsf{sr}(p), t))$

Axiom (a23) states, quite generally, that an inference p is a sampling bias (for agent i) when the objects of the relevant class cl are not present as observations in the source of p. A stronger alternative to definition (a23) states that any inference of any agent with that type of source, i.e., not including the required objects, results in a sampling bias, independently of the agent.

Expunging observations about a certain class of objects could be done on purpose, as we discussed, to balance normative or social requirements with statistical correlation.

5.4.2. Specification Bias

A specification bias ($b_{Sp_1,c,i}$) affects the relationship between the concepts used in the source and the concept c used in the target of an inference.

a24 $\mathsf{cf}(p, b_{Sp_1,c,i}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge \mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \forall o(\mathsf{cn}(o) = c_r \rightarrow \neg \mathsf{tp}(o, \mathsf{sr}(p), t))$

Definition (a24) simply states that an inference is a specification bias (for agent i) when the required concept c_r for c does not belong to the source of the inference. In this context, the relation $\mathsf{requires}(c, c_r)$ may take the meaning of the correct statistical correlation between c_r and c. So, by lacking c_r, the source of the inference is not based on the required evidence.

[18]Biases of this type may form a class of concepts $\mathsf{B}_K(x) \rightarrow \mathsf{B}(x)$, including all the concepts b such that, by replacing the c, c' and i, satisfy the condition of $\mathsf{cf}(p, b, t)$ in axiom (a22).

Other types of specification bias ($b_{Sp_2,c,i}$) can be defined, not using the requirement relation. We can say that an inference is biased if it is lacking an important class of concepts (cl') in its source (a25), or if it uses inappropriate concepts (from a certain class cl'') in its source (a26), or even both.

Moreover, analogous definitions may concern the target concept (a27), when c is a type of concept cl and the concepts used in the source are of an incompatible type cl'.

a25 $\mathsf{cf}(p, b_{Sp_1,c,i}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge$
$\mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \forall o(\mathsf{cf}(\mathsf{cn}(o), cl', t) \rightarrow \neg \mathsf{tp}(o, \mathsf{sr}(p), t))$

a26 $\mathsf{cf}(p, b_{Sp_2,c,i}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge$
$\mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \exists o(\mathsf{cf}(\mathsf{cn}(o), cl'', t) \wedge \mathsf{tp}(o, \mathsf{sr}(p), t))$

a27 $\mathsf{cf}(p, b_{Sp_3,c,i}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge$
$\mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \mathsf{cf}(c, cl, t) \wedge \forall o(\mathsf{tp}(o, \mathsf{sr}(p), t) \rightarrow \mathsf{cf}(\mathsf{cn}(o), cl', t))$

Stronger definitions of specification biases, again, make them independent of the agent. Notice also in this case that the absence of the correct statistical correlation, or of certain concepts, could be intentional, for instance when preventing the use of protected attributes in the dataset.

5.4.3. *Inductive Bias*

An inductive bias ($b_{I,c,i,e}$) is caused by the agent of an inference concerning a concept c and an entity e. We offer two general definitions, not entering the functioning of the agent i. An inference of agent i is biased according to $b_{I_1,c,i,e}$, see (a28), iff it is an inference whose target is a classification under concept c and even when i is provided with the correct evidence c_r to classify an entity e under c, i's inferences still classifies e under not c (i.e, $\bar{c}$). An inference of agent i is biased according to $b_{I_2,c,i,e}$, see (a29), iff it is an inference whose target is the classification of e under concept c and i has an inference p' with same source as p that classifies an entity e under $\bar{c}$, in a possibly different situation (here modelled by having the inference present at time t').

a28 $\mathsf{cf}(p, b_{I_1,c,i,e}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge \mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \mathsf{obj}(\mathsf{tr}(p)) = e \wedge$
$\exists p' o(\mathsf{pre}(p', t) \wedge \mathsf{ag}(p') = i \wedge \mathsf{cn}(o) = c_r \wedge \mathsf{tp}(o, \mathsf{sr}(p'), t) \wedge \mathsf{cn}(\mathsf{tr}(p')) = \bar{c} \wedge$
$\mathsf{obj}(\mathsf{tr}(p')) = e)$

a29 $\mathsf{cf}(p, b_{I_2,c,i,e}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge \mathsf{ag}(p) = i \wedge \mathsf{cn}(\mathsf{tr}(p)) = c \wedge \mathsf{obj}(\mathsf{tr}(p)) = e \wedge$
$\exists p' t'(\mathsf{pre}(p', t') \wedge \mathsf{ag}(p') = i \wedge \mathsf{sr}(p') = \mathsf{sr}(p) \wedge \mathsf{cn}(\mathsf{tr}(p')) = \bar{c} \wedge \mathsf{obj}(\mathsf{tr}(p')) = e)$

5.4.4. *Deployment Bias*

A way to approach, in general, a deployment bias ($b_{D,c,i}$) is to state that the objects used in the source are of a type that is not compatible with the object in the target. To represent compatibility, we use a concept cl such that all objects in the source belong to cl, whereas the object of the target classification belongs to an incompatible class $\bar{cl}$.

a30 $\mathsf{cf}(p, b_{D,c,i}, t) \leftrightarrow \mathsf{pre}(p,t) \wedge \mathsf{ag}(p) = i \wedge$
$\mathsf{cn}(\mathsf{tr}(p)) = c \wedge \mathsf{cf}(\mathsf{obj}(\mathsf{tr}(p)), \bar{cl}, t) \wedge \forall o(\mathsf{tp}(o, \mathsf{sr}(p), t) \rightarrow \mathsf{cf}(\mathsf{obj}(o), cl, t))$

An example could be a system of visual classification that has been trained on pictures of animals and deployed in a city centre to recognise vehicles.

6. Future Steps

In this paper, we have proposed a preliminary ontological representation of a general notion of bias, intended to comprehend a vast array of disciplines addressing biases that may affect both human and artificial agents. We have categorised and formalised biases as concepts classifying inferences, i.e. judgments produced by an agent using previously acquired observations. Inferences are thus meant to represent public utterances rather than mental states or internal calculations and, as we have seen, they are not necessarily correct.

We are aware that this work is only at its beginnings since we have limited our analysis to observations that have the form of classification acts, while not all observations display this nature: we may have, e.g., predictions or recommendations. Furthermore, with our framework, we have captured so far only a few types of biases among the plethora available in the literature.

Nonetheless, we are confident that the approach we have presented represents a novel direction of study on bias, one that can constitute a useful foundation for further studies and extensions of the proposed framework.

Acknowledgment: Research leading to this work was (partially) supported by: *BRIO Italian National Research Project PRIN 2020 2020SSKZ7R*.

References

[1] Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974;185(4157):1124-31.

[2] Korteling JE, Toet A. Cognitive biases. In: Sala SD, editor. Encyclopedia of Behavioural Neuroscience. 2nd ed. Elsevier; 2022. p. 610-9.

[3] Roulston K, Shelton SA. Reconceptualizing Bias in Teaching Qualitative Research Methods. Qualitative Inquiry. 2015;21(4):332-42. Available from: https://doi.org/10.1177/1077800414563803.

[4] Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable [Journal Article]. Annual Review of Sociology. 2014;40:31-53. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-soc-071913-043455.

[5] Zhou M, Abhishek V, et al.. Bias in Generative AI; 2024. arXiv preprint arXiv:2403.02726. Available from: https://arxiv.org/pdf/2403.02726.

[6] Russo M, Vidal ME. Leveraging Ontologies to Document Bias in Data. In: Calegari R, Dignum V, O'Sullivan B, editors. AEQUITAS 2024 Fairness and Bias in AI Proceedings of the 2nd Workshop on Fairness and Bias in AI co-located with 27th European Conference on Artificial Intelligence (ECAI 2024). vol. 3808. CEUR; 2024. .

[7] Reyero Lobo P, Daga E, et al. Semantic Web technologies and bias in artificial intelligence: A systematic literature review. Semantic Web. 2023 April;14(4):745-70. Available from: https://oro.open.ac.uk/85070/.

[8] Franklin JS, Bhanot K, et al. An Ontology for Fairness Metrics. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. AIES '22. New York, NY, USA: Association for Computing Machinery; 2022. p. 265–275. Available from: https://doi.org/10.1145/3514094.3534137.

[9] Brownstein M, Madva A. Implicit Bias. In: Zalta EN, Nodelman U, editors. The Stanford Encyclopedia of Philosophy. Spring 2025 ed. Metaphysics Research Lab, Stanford University; 2025. .

[10] Johnson GM. Algorithmic bias: on the implicit biases of social technology. Synthese. 2021;198(10):9941-61. Available from: https://doi.org/10.1007/s11229-020-02696-y.

[11] Johnson GM. Varieties of Bias. Philosophy Compass. 2024;19(7):e13011. Available from: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.13011.

[12] Gilovich T, Griffin D, et al., editors. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge: Cambridge University Press; 2002.

[13] Johnson GM. The Structure of Bias. Mind. 2020;129(516):1193-236.
[14] Kelly T. Bias, norms, introspection, and the bias blind spot. Philosophy and Phenomenological Research. 2024;108(1):81-105. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/phpr.12953.
[15] Brownstein M, Saul J. Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology. Oxford University Press; 2016. Available from: https://doi.org/10.1093/acprof:oso/9780198713241.001.0001.
[16] Danks D, London AJ. Algorithmic Bias in Autonomous Systems. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17; 2017. p. 4691-7. Available from: https://doi.org/10.24963/ijcai.2017/654.
[17] Fazelpour S, Danks D. Algorithmic Bias: Senses, Sources, Solutions. Philosophy Compass. 2021;16(8):e12760.
[18] Kordzadeh N, Ghasemaghaei M. Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems. 2022;31(3):388-409. Available from: https://doi.org/10.1080/0960085X.2021.1927212.
[19] González-Sendino R, Serrano E, et al. A Review of Bias and Fairness in Artificial Intelligence. International Journal of Interactive Multimedia and Artificial Intelligence. 2024 12/2024;9(1):5-17. Available from: https://www.ijimai.org/journal/bibcite/reference/3390.
[20] Hellström T, Dignum V, et al. Bias in machine learning - what is it good for? In: Saffiotti A, Serafini L, Lukowicz P, editors. NeHuAI@ECAI. vol. 2659 of CEUR Workshop Proceedings. CEUR-WS.org; 2020. p. 3-10. Available from: http://dblp.uni-trier.de/db/conf/ecai/nehuai2020.html#HellstromDB20.
[21] Mehrabi N, Morstatter F, et al. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv. 2021 Jul;54(6). Available from: https://doi.org/10.1145/3457607.
[22] Mandelbaum E. Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. Noûs. 2016;50(3):629-58. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12089.
[23] Mitchell C, De Houwer J, et al. The Propositional Nature of Human Associative Learning. The Behavioral and brain sciences. 2009 05;32:183-98; discussion 198.
[24] Gendler T. Alief and Belief. The Journal of Philosophy. 2008 10;105:634-63.
[25] Petty RE, Briñol P, et al. The Meta–Cognitive Model (MCM) of Attitudes: Implications for Attitude Measurement, Change, and Strength. Social Cognition. 2007;25(5):657-86. Available from: https://doi.org/10.1521/soco.2007.25.5.657.
[26] Gawronski B, Brannon SM. Attitudes and the Implicit-Explicit Dualism. In: Albarracín D, Johnson BT, editors. The handbook of attitudes: Volume 1: Basic principles. 2nd ed. New York: Routledge; 2017. p. 158-96.
[27] Rojszczak A. From the Act of Judging to the Sentence: The Problem of Truth Bearers From Bolzano to Tarski. Springer; 2005.
[28] Austin JL. How to do things with words. Massachusetts: Harvard University Press; 1962. Edited by J.O. Urmson and Marina Sbisà.
[29] Masolo C, Borgo S, et al. WonderWeb Deliverable D18. CNR; 2003.
[30] Borgo S, Ferrario R, et al. DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering. Applied Ontology. 2022 Jan;17(1):45-69.
[31] Masolo C, Vieu L, et al. Social Roles and their Descriptions. In: KR; 2004. p. 267-77.
[32] Masolo C, Botti Benevides A, et al. The interplay between models and observations. Applied Ontology. 2018;13(1):41-71.
[33] Kahneman D. Thinking, Fast and Slow. New York, USA: Farrar, Straus and Giroux; 2011.
[34] Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981;211(4481):453-8. Available from: http://www.sciencemag.org/cgi/content/abstract/211/4481/453.
[35] Block LG, Keller PA. When to Accentuate the Negative: The Effects of Perceived Efficacy and Message Framing on Intentions to Perform a Health-Related Behavior. Journal of Marketing Research. 1995;32(2):192-203. Available from: https://doi.org/10.1177/002224379503200206.