Toward A Causal Framework for Modeling Perception

Jose M. Alvarez[1], **Salvatore Ruggieri**[2]

[1]Department of Computer Science, KU Leuven
[2]Department of Computer Science, University of Pisa
josemanuel.alvarez@kuleuven.be, salvatore.ruggieri@unipi.it

Abstract

Perception occurs when individuals interpret the same information differently. It is a known cognitive phenomenon with implications for bias in human decision-making. Perception, however, remains understudied in machine learning (ML). This is problematic as modern decision flows, whether partially or fully automated by ML applications, always involve human experts. For instance, how might we account for cases in which two experts interpret differently the same deferred instance or explanation from a ML model? Addressing this and similar questions requires first a formulation of perception, particularly, in a manner that integrates with ML-enabled decision flows. In this work, we present a first approach to modeling perception causally. We define perception under causal reasoning using structural causal models (SCMs). Our approach formalizes individual experience as additional causal knowledge that comes with and is used by the expert decision-maker in the form of a SCM. We define two kinds of probabilistic causal perception: structural and parametrical. We showcase our framework through a series of examples of modern decision flows. We also emphasize the importance of addressing perception in fair ML, discussing relevant fairness implications and possible applications.

1 Introduction

The same piece of information can be interpreted differently by individuals. The late Kahneman (2021) refers to this cognitive phenomenon as perception. It is a product of the mental heuristics humans use to enable fast decision-making under uncertainty. These heuristics and, in turn, perception are shaped by individual experience, which can lead to biased decision-making. The Linda Problem (Tversky and Kahneman 1974, 1981, 1983), for instance, illustrates how perception overrides logical reasoning when individuals rely on intuitive judgment. In a series of famous experiments, participants committed the conjunction fallacy by judging the conjunction ("Linda is a bank teller and a feminist") as more probable than a single constituent ("Linda is a bank teller"), violating the basic laws of probability that govern rational decision-making. These and similar decision-making experiments (Kahneman 2011; Kahneman et al. 2016; Thaler and Sunstein 2008), together with the growing regulatory emphasis on trustworthy machine learning (ML) applications (Álvarez et al. 2024), underscore the need to consider perception as a core problem in ML.

Modern decision processes are powered by ML applications and always involve some degree of human decision-making (Ruggieri et al. 2023; Scantamburlo, Baumann, and Heitz 2024; Ruggieri and Pugnana 2025). In learning to defer (LtD), for example, the goal is to learn a ML model that abstains from predicting on instances it is not certain of and defers the decision to a human expert (Madras, Pitassi, and Zemel 2018). In explainable artificial intelligence (xAI), for example, the goal is to develop methods that explain to an expert the predictions of a ML model (Guidotti et al. 2019). Perception can clearly occur in both settings when multiple individual experts are involved, potentially impacting the outputs of these ML applications.

How might we develop reliable LtD and xAI methods when experts may disagree on the same deferred instance or explanation? These questions are often ignored when building modern decision flows (see, e.g., Cabitza, Campagner, and Basile (2023); Fahimi et al. (2024); Srivastava, Heidari, and Krause (2019)). Importantly, addressing this gap requires a formalization of perception that aligns with the standard ML problem formulation, which is probabilistic (Pearl 1988, 2009; Goodfellow et al. 2016). In this work, we provide a first approach to this objective.

1.1 Our Contributions

How can we formally represent individuals who disagree in their interpretation of the same information? We first propose a probabilistic framework for perception in which individuals construct disagreeing probability distributions, $P(\mathbf{X})$, for the same information $\mathbf{X}$. Under this framework, for example, we can explain why one expert assigns a higher probability to a positive classification for a deferred instance compared to another expert. We then propose causal perception by extending the initial formulation to account for probabilistic causal reasoning. Here, individuals are equipped with structural causal models (SCMs) (Pearl 2009) that not only describe $P(\mathbf{X})$ but also govern how they reason about hypothetical scenarios. As a result, individuals may disagree not only on observed distributions but also on interventional and counterfactual ones. Under this framework, we can describe how the two experts arrive at different classification probabilities by answering counterfactual questions, e.g., on

a deferred instance. Our framework draws on foundational work in probabilistic and causal reasoning under uncertainty (Pearl 1988, 2009). We adopt a particular view of perception as disagreement in the interpretation of information conveyed through probabilities.

We define two kinds of causal perception: structural and parametrical. In structural perception individuals disagree on the cause-effect relationships, while in parametrical perception individual disagree on the causal effect between an established cause-effect pair. We complement these two kinds of causal perception by providing an initial approach to model the processes by which individuals construct their SCM. Further, we discuss areas of interest within fair ML and point at promising directions of future work involving perception, such as situated bias and reformulating sensitive attributes as loaded attributes. To showcase our framework, we present a series of examples involving ML applications and individual experts.

In the rest of the paper, we introduce the necessary background in Section 2. We then define probabilistic perception, or perception, in Section 3 and causal probabilistic perception, or causal perception, in Section 4. We present a first approach to the construction of the individual SCMs behind causal perception in Section 5. We discuss the fair ML implications in Section 6. We conclude in Section 7.

1.2 Related Work

Perception has been studied mainly by psychologist, with views on it varying within the field. We view perception as individual differences in interpreting the same information shaped by mental heuristics and expressed as probabilities; we discuss our choice later in Section 3. The focus within such view has been on the representativeness heuristic, in which an event is made to be more representative of a class than what it actually is, as measured by a higher probability (Tversky and Kahneman 1974, 1981, 1983). Bayesian modeling—in which the baseline representativeness of an event (the prior) is adjusted by the agent based on her experience (the posterior)—remains the common approach for modeling explicitly the representativeness heuristic and, thus, implicitly perception (Costello 2009; Bordalo et al. 2016; Tentori 2022). Such view belongs to a cognitive-bias-driven interpretation of perception in which perception influences rational decision-making under uncertainty (Kahneman 2011). There are, for instance, constructivist (Gregory 1970) and sensorial (Goldstein 1996) views that study the interpretative aspect of perception. We are the first to use SCMs under this view of perception, which allows to frame it also in terms of hypothetical distributions that represent counterfactual reasoning. Further, we extend previous works since SCMs, or causal Bayesian networks (Pearl 2009), include Bayesian reasoning. The proposed causal perception framework, thus, can be viewed as a comparison of two individuals' posterior judgments and the causal theories that underlie those judgments.

ML researchers are increasingly interested in cognitive biases (see, e.g., Bengio (2019); Booch et al. (2021)). Such works focus on how to create ML-driven intelligent systems that improve over and potentially replace the irrational human expert. It is an active and interdisciplinary line of research (see, e.g., Agrawal, Gans, and Goldfarb (2019)). Our work adds to this growing line of research by making use of causal reasoning as a basis for modeling human-like reasoning (Schölkopf 2022; Pearl 2009). Further, there are narrower lines of ML work that study these cognitive biases and their impact on the human user of the ML application. We highlight those studying the human-in-the-loop problem from a fairness and accuracy perspective (see, e.g., De et al. (2020); Mozannar et al. (2023); Palomba et al. (2025)), which includes LtD (Madras, Pitassi, and Zemel 2018) and xAI (Rong et al. 2024). Such works consider the human's interaction with the ML model, incorporating it within the problem formulation. With some exceptions (Caraban and Karapanos 2020; Rastogi et al. 2022; Yang, Folke, and Shafto 2025), these works, however, do not give agency to the human user and formulate it as an additional and costly decision-maker, ignoring any influence from individual experience. Different from these works, we explore the setting of having multiple users that disagree on the information provided. In that sense, our work gives more agency to the user by recognizing the role of perception among a set of heterogeneous users.

The role of these cognitive biases in fair ML is largely unexplored, with some exceptions. Some of these works show how the cognitive biases can be exacerbated by the xAI techniques used by the expert to explain a model (Bertrand et al. 2022) and can affect the expert's evaluation of a model's output (Echterhoff, Yarmand, and McAuley 2022). Further, given the problem of context-aware fairness, where we recognize that fairness has different meanings across humans, works like Srivastava, Heidari, and Krause (2019) and Yaghini, Krause, and Heidari (2021) design user experiments to test for the human perception of fairness. These experiments show that the fairness of an outcome can be judged differently depending on who is the individual judging. We add to these works by formalizing perception itself through causal reasoning. In Section 5.3 we further discuss perception within the causal fair ML literature.

1.3 A Running Example

Throughout the paper we use a series of ML-enabled decision flow scenarios based on Example 1.

Example 1 *(College Admissions) An admissions officer (the decision maker or DM) chooses the incoming class based on the applicants' profiles. Assume a decision flow in which the officer admits, $Y = 1$, or rejects, $Y = 0$, applicants based on their SAT results, X_1, high-school GPA, X_2, and suitability scores $f(X_1, X_2) = G \in [0, 1]$ where f is a ML model trained by the college. The officer, through the applicants' motivation letters, also has access to their address, Z. Consider three scenarios in which the officer relies on f with varying degrees for the decision:*
(i) f abstains from classifying an applicant on which it is not confident and the officer must classify this applicant;
(ii) f provides the same score for two applicants and the officer must choose one between these two applicants;
(iii) f alone derives Y using G to rank applicants and ad-

mits the top-k ones where k is set by the officer.

The above example is an extended version of Kleinberg et al. (2019)'s college admissions tiebreaker example. Scenarios *(i)* and *(ii)* represent a partially automated decision flow in which f aids the DM, while scenario *(iii)* represents a fully automated decision flow in which f replaces the DM. These are high-level but common ML-enabled scenarios. Additional context is provided for each scenario as we extend Example 1 moving forward.

2 Background

In this work, we represent information as a set of p discrete random variables $\mathbf{X} = X_1, \ldots, X_p$. Let $P(\mathbf{X})$ denote the joint probability distribution of $\mathbf{X}$ with $P(\mathbf{x})$ representing the joint probability that $\mathbf{X}$ equals the p instances $\mathbf{x} = x_1, \ldots, x_p$. We treat individuals, like a human user, and objects, like a ML application, as agents.

Senders and Receivers. Borrowing from signaling games (Spence 1973), we consider two types of agents. A *sender* $S \in \mathcal{S}$, with $\mathcal{S}$ denoting the set of senders, is an agent that provides information while a *receiver* $R \in \mathcal{R}$, with $\mathcal{R}$ denoting the set of receivers, is an agent that interprets the provided information. We only borrow this distinction as our agents are non-strategic and engage in a single one-shot game, meaning that the sender sends and the receiver receives information once without tricking one another. For an overview of signaling games, see Sobel (2020).

Structural Causal Models. A SCM (Pearl 2009) describes the data-generating model of a process, allowing to reason how the p random variables in $\mathbf{X}$ relate to each other as cause-effect pairs and how these pairs determine $P(\mathbf{X})$. Formally, a SCM $\mathcal{M}$ is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{X}, \mathbf{F} \rangle$ that transforms a set of p latent variables $\mathbf{U} \sim P(\mathbf{U})$ into a set of p observed variables $\mathbf{X}$ according to a set of structural equations $\mathbf{F}$ such that:

$$P(\mathbf{U}) = P(U_1, \ldots, U_j) \quad X_j := f_j(X_{pa(j)}, U_j) \tag{1}$$

for $j = 1, \ldots, p$ where $U_j \in \mathbf{U}$, $X_j \in \mathbf{X}$, and $f_j \in \mathbf{F}$. Each function f_j maps the latent variable U_j to the observed variable X_j based on the subset of observed variables that directly cause X_j, or its causal parents $X_{pa(j)}$.

The parental relations in a SCM induce a *causal graph* $\mathcal{G}$, in which the nodes represent random variables and the directed edges between them causal relations. We focus on acyclic graphs, meaning there are no loops in $\mathcal{G}$. This assumption turns $\mathcal{G}$ into a *directed acyclical graph* (DAG) and ensures that information does not travel backwards, which is a common assumption (Binkyte et al. 2022).

Linear SCMs. We focus on the *additive noise models* (ANMs), where f_j is a linear transformation:

$$f_j := \sum_{i=1}^{|pa(j)|} \beta_{ij} \cdot X_{pa(j)_i} + U_j \tag{2}$$

with $\beta_{ij} \in \mathbb{R}$ denoting the *causal weight* of the i-th parent $pa(j)_i$ of X_j. Our discussion applies to all SCMs, but we consider the ANM case for readability.

Causal Reasoning and Its Implied Distributions. The SCM $\mathcal{M}$ allows to reason about $P(\mathbf{X})$ in terms of observed and hypothetical scenarios.[1] For the observed scenario, or *what is*, it is possible to disentangle the joint probability distribution $P(\mathbf{X})$ by factorizing it as a product of cause-effect pairs given the SCM $\mathcal{M}$:

$$P(\mathbf{X}) = \prod_{i=1}^{p} P\Big(X_i \mid X_{pa(i)}\Big) \tag{3}$$

which simplifies reasoning about $P(\mathbf{X})$, as it states that X_i is conditionally independent of its non-descendants given its parents $X_{pa(i)}$. This property is known as the *Markovian condition* (Peters, Janzing, and Schölkopf 2017).

For the hypothetical scenarios, or *what if*, it is possible to generate new distributions of $P(\mathbf{X})$ by intervening the SCM $\mathcal{M}$. This is because SCMs build on an interventionist account of causality (Woodward 2005). An intervention on a single variable X_i is done via the *do-operator*, $do(X_i = x_i)$, which replaces the structural equation in $\mathbf{F}$ for the variable X_i with the value x_i. Interventions apply similarly for multiple variables, $do(X_i = x_i, X_j = x_j)$, replacing the structural equations for each variable individually. Let $\mathcal{I}_\mathbf{X}$ denote the *set of all interventions*, which is an index set with each index representing a specific intervention on the variables $\mathbf{X}$. We use $\emptyset \in \mathcal{I}_\mathbf{X}$ to denote the null intervention. As Rubenstein et al. (2017) demonstrate, $\mathcal{I}_\mathbf{X}$ has a *natural partial ordering*, in which for interventions $i, j \in \mathcal{I}_\mathbf{X}$, $i \leq_\mathbf{X} j$ if and only if i intervenes on a subset of the variables that j intervenes on and sets them equal to the same values as j. For instance, $do(X_i = x_i) \leq_\mathbf{X} do(X_i = x_i, X_j = x_j)$. It means the j intervention can be done after the i intervention without needing to change the modifications done by the i intervention on the SCM $\mathcal{M}$.

Each intervention implies a well-defined joint distribution of $\mathbf{X}$ variables $P(\mathbf{X})^{do(i)}$ for $i \in \mathcal{I}_\mathbf{X}$, called the i-interventional distribution. Following Rubenstein et al. (2017), we define the *poset of all interventional distributions implied* by the SCM $\mathcal{M}$, where $\leq_\mathbf{X}$ is the natural partial ordering inherited from $\mathcal{I}_\mathbf{X}$, as:

$$\mathcal{P}_\mathbf{X} := \Big(\Big\{ P(\mathbf{X})^{do(i)} : i \in \mathcal{I}_\mathbf{X} \Big\}, \leq_\mathbf{X} \Big). \tag{4}$$

By definition, $P(\mathbf{X}) \in \mathcal{P}_\mathbf{X}$. Further, $\mathcal{P}_\mathbf{X}$ is a singleton comprised of $P(\mathbf{X})$ when $\mathcal{I}_\mathbf{X} = \{\emptyset\}$. Furthermore, intuitively, $\mathcal{P}_\mathbf{X}$ represents all possible ways of reasoning causally about variables $\mathbf{X}$ as implied by a SCM $\mathcal{M}$.

3 Perception of Probability

We first present perception in its most basic form. The goal is to formulate when two individuals, the receivers, interpret differently the same information from a ML application, formalized as the sender. We consider, in particular, the setting in which these individuals act as some sort of decision maker within a decision flow.

[1] Pearl and Mackenzie (2018) present three levels of reasoning: observational (*what is*), interventional (*what if*), and counterfactual (*what would have been if*). The latter two represent the hypothetical scenario. We use a simpler distinction.

We deal with perception of probability as we assume that the individuals manifest their interpretations of the information received using probabilities. This assumption is rooted in how humans reason under uncertainty, which has a long and established tradition across multiple fields that study bias in human decision-making (see, e.g., Kahneman (2011); Thaler and Sunstein (2008)). We find this to be a reasonable assumption. Whether it is because the setting is inherently stochastic or the individual cannot possibly observe all the variables involved, there is always some degree of uncertainty involved in decision flows (Pearl 1988). Probabilistic reasoning helps capture the uncertainty behind making a choice based on a ML application's outcome. Further, probability theory is one of the pillars of ML (Goodfellow et al. 2016). By describing perception in terms of probabilities, we can relate it to standard ML problem formulations. Let us now offer a preliminary definition of perception of probability entirely based on this intuition.

Definition 3.1 *(Perception) For receivers $R_i, R_j \in \mathcal{R}$, given the information by sender $S \in \mathcal{S}$ in the form of a random variable X, perception occurs when, for a threshold $\epsilon \in \mathbb{R}^+$, we have that:*

$$d\left(P_{R_i}(X), P_{R_j}(X)\right) > \epsilon \tag{5}$$

where the probability distributions $P_{R_i}(X)$ and $P_{R_j}(X)$ represent the interpretations by R_i and R_j, respectively, of the information X, and $d(\cdot,\cdot)$ denotes a suitable distance metric between probability distributions.

Remark 3.1 *(From variable to variables) Definition 3.1 naturally applies to the p random variables $\mathbf{X}$. In that case, $P_{R_i}(\mathbf{X})$ and $P_{R_j}(\mathbf{X})$ represent the joint probability distributions of the receivers for information $\mathbf{X}$. Further, we can consider the particular case in which one of the random variables denotes the target Y and $P_{R_i}(Y|\mathbf{X})$ and $P_{R_j}(Y|\mathbf{X})$ represent the conditional joint probability distributions of the receivers for information $\mathbf{X}$ with $p-1$.*

Remark 3.2 *(Instance-based perception) Implicit to Definition 3.1 is a disagreement at the instance level $X = x$, meaning $d\big(P_{R_i}(x), P_{R_j}(x)\big) > \epsilon$ where $P_{R_i}(x)$ and $P_{R_i}(x)$ are probabilities and $d(\cdot,\cdot)$ a suitable distance metric.*[2] *Intuitively, if a receiver can construct the probability distribution $P(X)$, then it can also compute the probability of any instance $x \in X$. It is also the case for the joint probabilities of the p instances in $\mathbf{x}$ of the p random variables in $\mathbf{X}$.*

With Definition 3.1, we are interested in capturing how perception—as a disagreement in probability—would look like formally. We take for granted the process behind how these two receivers construct(ed) $P(X)$ and focus on the fact that there is a probabilistic disagreement between them. The fact that there is disagreement of probability between R_i and R_j only has an impact when these two receivers must make a decision within the same decision-making context based on $P(X)$. If $d\big(P_{R_i}(X), P_{R_j}(X)\big) \leq \epsilon$, then we would not have a disagreement and we would be indifferent between R_i and R_j. In that case, these two receivers, for our purposes, would be indistinguishable within said context. Let us consider the next example.

Example 2 *Let us consider Example 1, scenario (i). Suppose two admissions officers, R_1 and R_2, receive the same deferred applicant by the abstaining ML model f with profile $\mathbf{x} = \langle x_1, x_2 \rangle$. Suppose that perception occurs as each officer constructs different $P(Y|X_1, X_2)$, such that $P_{R_1}(Y|x_1, x_2) > P_{R_2}(Y|x_1, x_2)$, for which officer R_1 accepts while R_2 rejects the applicant.*

As Example 2 illustrates, Definition 3.1 dwells exclusively with the observational probability distribution or the *what is*. It does not allow for hypothetical scenarios, and the applicant is judged based on what is observed in his or her profile. There is no, for instance, counterfactual reasoning involved when deciding over the deferred applicant. We expand such a setting in Section 4.

3.1 The Representativeness Heuristic

Although Definition 3.1 cares only for the disagreement in probabilistic reasoning, we briefly discuss the mental heuristic commonly associated with perception as we view it in this work. We do so to further support our focus on perception of probability. The representativeness heuristic is one among many heuristics used by humans that lead to biased decision-making under uncertainty (Tversky and Kahneman 1974). Other heuristics, however, such as availability and anchoring, could also motivate our definition of perception.

Probabilities can be understood as quantifying *how representative* an instance x is of a random variable X. Under this premise, for instance, Definition 3.1 reflects the setting in which two individuals can judge differently the representativeness of the same instance x relative to the random variable X. Similarly, the *representativeness heuristic* is used when humans assess how much one instance x resembles another instance x' that is known to belong to X, which may distort the estimation of $P(x)$. This is because, in such a case, reasoning centers on *resemblance between instances*. The more one instance resembles the other instance, the more representative the former is of the latter instance. The question "What is the probability that x belongs to the random variable X?" becomes "To what degree the instance x resembles the other known instance x' in the random variable X?". If the resemblance is high, then we judge the probability $P(x)$—read as x belongs to X or, equivalently, x is generated by X—to be high. Here, it helps overall to reason in terms of *degree of representativeness*, which translates naturally into how we understand probabilities.

Definition 3.2 *(Degree of Representativeness) For the random variable X with known instance x', the probability $P(x)$ is estimated by the degree of representativeness of x as an instance of X based on its resemblance to x'. Given a distance $d(\cdot,\cdot)$ between instances, we have $P(x) \approx P(x')$ as $d(x, x') \approx 0$. This definition extends to the joint probability distribution of the collection of random variables $\mathbf{X}$, with instances $\mathbf{x}$ and known instances $\mathbf{x}'$.*

[2] Here, the focus is on the specific instance x as, indeed, distance among probability distributions in Equation 5 may be $\leq \epsilon$, with still a number of instances having distance $> \epsilon$.

It follows from Definition 3.2 that, if the known instance x' is viewed as *representative of* X, denoted by a high-enough $P(x')$ within the relevant context, then x is also representative of X as captured by $P(x)$. The degree of representativeness is clearly a function of what someone considers representative of X as captured by x' and $P(x')$. With respect to our framework for perception, it implies that receivers have their own x' that they use to construct $P(x)$.

Example 3 *The occurrence of perception in Example 2 is driven by each admissions officer resorting to their own degree of representativeness when evaluating the deferred applicant with profile* $\mathbf{x} = \langle x_1, x_2 \rangle$*. Underlying the competing* $P_{R_1}(Y|x_1, x_2)$ *and* $P_{R_2}(Y|x_1, x_2)$ *is that each officer compares* $\mathbf{x}$ *relative to their own* $\mathbf{x}'$*.*

For our purposes, we view this process as representing *individual experience*. It is private information known only to the receivers. How such individual experience is constructed over time and elicited while reasoning by the receiver is beyond the scope of this work. We do, however, highlight another of Kahneman's work: *norm theory* (Kahneman and Miller 1986), which theorizes how individuals respond to the familiarity of an event by recruiting and creating alternative scenarios. Under this theory, R recruits a number of representations about an event. These representations are based on what R views as a normal, with each scenario having a set of elements and each element having a set of features. These representations can be aggregated into a single scaled representation denoting the most common alternative, or *the norm*, according to R. Such a process could help formulate under a ML approach how individual experience is constructed, stored, and used by the receiver. We leave this for future work.

4 Formulating Causal Perception

We now extend Definition 3.1 to account for probabilistic causal reasoning. We want to capture disagreement between receivers beyond the *what is* probability distribution by also considering the *what if* probability distributions of the information sent by the sender. We rely on a SCM $\mathcal{M}$ to model the causal generative distribution $P(\mathbf{X})$. Notably, the single random variable case $P(X)$ is of no interest here as it lacks a parent or child to reason about.

4.1 Why Causality?

Our modeling choice is based on three factors. First, under the premise that humans use probabilistic reasoning for decision-making under uncertainty (Kahneman 2011) as argued in Section 3, the properties of a SCM $\mathcal{M}$ (Equation 1) offer a structured way to describe the information $\mathbf{X}$ as represented by $P(\mathbf{X})$. A SCM $\mathcal{M}$ allows to factorize $P(\mathbf{X})$ (Equation 3) and to reason under hypothetical scenarios about $P(\mathbf{X})$ (Equation 4), in principle, similar to how humans interpret information and accumulate knowledge (Woodward 2005; Pearl 2009; Pearl and Mackenzie 2018). Second, SCMs are increasingly used by ML researchers to approximate human-like reasoning (Schölkopf 2022; Schölkopf et al. 2021). Given this trend, our proposed framework is compatible with the next wave of ML applications (see, e.g., van Steenkiste et al. (2019); Dittadi et al. (2021)). Third, a SCM, in particular through its DAG $\mathcal{G}$, is a useful tools for engaging multiple stakeholders. A DAG, for instance, can be used to draw the assumptions about $\mathbf{X}$ and its data generating model in sensitive settings like discrimination testing (Álvarez and Ruggieri 2023, 2025) and synthetic data generation (Baumann et al. 2023). The DAG can essentially "draw" relevant individual experience.

We do not view SCMs as equivalent to human reasoning. Rather, based on these three factors, we simply view SCMs as a useful and pragmatic tool for formalizing human reasoning about $\mathbf{X}$ in terms of probabilities (similar, e.g., to Loftus (2024)). We are aware that this is not a view widely held within the fairness community (see, e.g., Hu and Kohler-Hausmann (2020)). Future work could explore non-causal approaches to extend Definition 3.1 that also capture reasoning about hypothetical scenarios.

4.2 Perception of Implied Probabilities

Definition 4.1 *(Causal Perception) For receivers* $R_i, R_j \in \mathcal{R}$ *with SCM* $\mathcal{M}_{R_i}$ *and* $\mathcal{M}_{R_j}$ *for the information* $\mathbf{X}$ *provided by sender* $S \in \mathcal{S}$*, causal perception occurs when for a threshold* $\epsilon \in \mathbb{R}^+$*:*

$$\overline{d}\left(\mathcal{P}_{\mathbf{X}_{R_i}}, \mathcal{P}_{\mathbf{X}_{R_j}}\right) > \epsilon \tag{6}$$

where $\mathcal{P}_{\mathbf{X}_{R_i}}$ *and* $\mathcal{P}_{\mathbf{X}_{R_j}}$ *represent the poset of all distributions of information* $\mathbf{X}$ *induced by* $\mathcal{M}_{R_i}$ *and* $\mathcal{M}_{R_j}$*, and* $\overline{d}(\cdot, \cdot)$ *denotes a suitable aggregated distance measure between two sets of probability distributions.*

The Definition 4.1 extends Definition 3.1 into the realm of probabilistic causal reasoning. It defines perception beyond a disagreement on the representation of $\mathbf{X}$ in terms of probabilities by also accounting for *any* disagreement on the interventional distributions of the causal models. In Example 1, for instance, we would go from disagreeing on "What is the probability of an applicant being successful?" to also disagreeing on "What is the probability of an applicant being successful had the applicant been different on one of its variables?". The latter kind of disagreement is inherently linked to the SCM of each receiver.

For Equation (6), we propose in practice to compute the distance as an aggregation of these distributions, such as an average or a maximum (see, e.g., Goldenberg and Webb (2019)). Formally, given a set of interventions $\mathcal{I}_{\mathbf{X}}$, we consider the *aggregated perception* for R_i and R_j in $\mathcal{R}$ as:

$$\bar{d}(\mathcal{P}_{\mathbf{X}_{R_i}}, \mathcal{P}_{\mathbf{X}_{R_j}}) = \Phi_{l \in \mathcal{I}_{\mathbf{X}}} d\left(P_{R_i}^{do(l)}(\mathbf{X}), P_{R_j}^{do(l)}(\mathbf{X})\right) \tag{7}$$

such that $P_{R_i}^{do(l)}(\mathbf{X})$ denotes the *l*-interventional distribution in the SCM $\mathcal{M}_{R_i}$ of receiver R_i after performing the *l*-intervention in $\mathcal{I}_{\mathbf{X}}$. The same applies for receiver R_j and its SCM $\mathcal{M}_{R_j}$. The choice of the distance function $d(\cdot, \cdot)$ and the aggregation function Φ in Equation (7) will depend on the context and the kind of disagreement we wish to capture. We could, for instance, define $d(\cdot, \cdot)$ as the Kullback–Leibler divergence and Φ as the average.

Remark 4.1 *(Specific interventional distributions) Notably, with Equation 7 we can also move away from "any disagreement" to "a specific disagreement" in terms of interventional distributions between receivers. In the special case in which no intervention occurs, or $\mathcal{I}_{\mathbf{X}} = \{\emptyset\}$, Equation* (7) *boils down to Equation* (5)*.*

Following up the previous remark, in the case of a specific l-intervention, or $\mathcal{I}_{\mathbf{X}} = \{l\}$, Definition 4.1 becomes:

$$d\left(P_{R_i}(X)^{do(l)}, P_{R_j}(X)^{do(l)}\right) > \epsilon. \tag{8}$$

Equation (8) highlights how Definitions 3.1 and 4.1 relate to each other as the difference in, respectively, the observational and the l-interventional distributions. Here, we study the disagreement of the interventional distribution resulting from the l-intervention. We can generate these interventional distributions via the SCMs $\mathcal{M}_{R_i}$ and $\mathcal{M}_{R_j}$.

4.3 Structural and Parametrical Perception

Central to Definition 4.1 are the SCMs of each receiver for $P(\mathbf{X})$. If we assume that receivers draw all variables $\mathbf{X}$ using a SCM $\mathcal{M}$, then two kinds of causal perception emerge: structural and parametrical. This assumption implies that the receivers, who receive the same information $\mathbf{X}$ from the sender, use all elements in $\mathbf{X}$, meaning each $X \in \mathbf{X}$ is also a variable in the SCM $\mathcal{M}$. We consider the particular case of ANM (Equation 2) for illustrative purposes.

Definition 4.2 *(Structural Causal Perception) It occurs when receivers $R_i, R_j \in \mathcal{R}$ disagree on the cause-effect pairs in, respectively, $\mathcal{M}_{R_i}$ and $\mathcal{M}_{R_j}$ for $P(\mathbf{X})$. It implies distinct DAGs $\mathcal{G}_{R_i}$ and $\mathcal{G}_{R_j}$.*

Definition 4.3 *(Parametrical Causal Perception) It occurs when receivers $R_i, R_j \in \mathcal{R}$ agree on the cause-effect pairs for $P(\mathbf{X})$, meaning the their DAGs $\mathcal{G}_{R_i}$ and $\mathcal{G}_{R_j}$ are the same, but disagree on the sets of structural equations $\mathbf{F}$ in respectively, $\mathcal{M}_{R_i}$ and $\mathcal{M}_{R_j}$.*

For instance, let $\mathbf{X} = \{X_1, X_2\}$. In structural causal perception, R_i draws $X_1 \rightarrow X_2$ while R_j draws $X_2 \rightarrow X_1$. The distinct DAGs imply structurally different equations in the SCM as cause-effect pairs are reversed. In parametrical causal perception, assume R_i and R_j agree on $X_1 \rightarrow X_2$. Broadly, it means both will have a structural equation of the form $X_2 := \beta \times X_1 + U$ in their SCM. However, for R_i $\beta > 0$ while for R_j $\beta < 0$. These two equations are parametrized differently coefficient-wise.

Example 4 *Let us consider Example 1, scenario (ii). Suppose two admissions officers, R_1 and R_2, are tasked to choose independently which of the two applicants tied by the score G from the ML model f is admitted. Two additional points to consider. First, the fact that f gives the same score to both applicants implies comparable profiles, which can be summarized by a single $\mathbf{x} = \langle x_1, x_2 \rangle$. These two applicants are indistinguishable to f. Second, this fact allows the officers to derive the decision Y without G. They can now consider other aspects of the applicants' profiles. Further suppose the officers resort to the applicants' zip code Z for the tiebreaker. In particular, both officers reason about $P(Y|X_1, X_2, Z)$ by asking "What would have been the score of an applicant had she or he been from the $Z = z$ zip code?" We can represent such hypothetical question using the $do(Z := z)$ on the SCMs $\mathcal{M}_{R_1}$ and $\mathcal{M}_{R_2}$ of each officer, thus, looking at the interventional joint probability distribution that is $P(Y|X_1, X_2, do(Z := z))$.*

Let us first consider the case of structural perception. Suppose two competing DAGs for $\{Y, X_1, X_2, Z\}$: let $G_{R_1} = \{Z \rightarrow X_1, Z \rightarrow X_2, X_2 \rightarrow X_1, X_1 \rightarrow Y, X_2 \rightarrow Y\}$ and $G_{R_2} = \{Z \rightarrow X_2, X_2 \rightarrow X_1, X_1 \rightarrow Y, X_2 \rightarrow Y\}$, meaning R_2 does not agree with R_1 that $Z \rightarrow X_1$. Given these DAGs, each office factorizes the interventional joint probability distribution of interest differently when answering the what if question. We have $P(Y|X_1, X_2)P(X_2|Z)P(X_1|X_2, Z)P(Z)$ under $\mathcal{G}_{R_1}$ and $P(Y|X_1, X_2)P(X_2|Z)P(X_1|X_2)P(Z)$ under $\mathcal{G}_{R_2}$. The following set of structural equations corresponds to $\mathcal{G}_{R_1}$:

$$Z := U_1 \tag{9}$$

$$X_1 := \beta_1 \cdot X_2 + \beta_2 \cdot Z + U_2 \tag{10}$$

$$X_2 := \beta_3 \cdot Z + U_3 \tag{11}$$

$$Y := \beta_4 \cdot X_1 + \beta_5 \cdot X_2 + U_4 \tag{12}$$

and a similar set of structural equations corresponds to $\mathcal{G}_{R_2}$ with the key difference that (10) becomes:

$$X_1 := \beta_1 \cdot X_2 + U_2 \tag{13}$$

as we exclude the term $\beta_2 \cdot Z$ since there is no $Z \rightarrow X_1$ cause-effect pair in R_2's DAG. When reasoning about Z and its impact on the other variables, each officer considers distinct interventional distributions due to the different factorizations of $P(\mathbf{X})$ as well as the competing set of structural equations. R_1 constructs $P(Y|X_1, X_2)P(X_2|z)P(X_1|X_2, z)P(z)$ while R_2 $P(Y|X_1, X_2)P(X_2|z)P(X_1|X_2)P(z)$. Similarly, in terms of the set of structural equations, when substituting each Z for the value z using the do-operator for R_1 we obtain:

$$Y := \beta_4 \cdot (\beta_1 \cdot X_2 + \beta_2 \cdot z + U_2) + \beta_5 \cdot (\beta_3 \cdot z + U_3) + U_4 \tag{14}$$

while for R_2 we obtain:

$$Y := \beta_4 \cdot (\beta_1 \cdot X_2 + U_2) + \beta_5 \cdot (\beta_3 \cdot z + U_3) + U_4 \tag{15}$$

where the term $\beta_4 \cdot \beta_2 \cdot z$ is missing for R_2 with respect to R_1 when making the decision. Such a difference can lead to a different applicant chosen by each officer based on Z.

Let us now consider the case of parametrical perception. Suppose R_1 and R_2 agree on the previous DAG $\mathcal{G}_{R_1}$. However, for equation (10) suppose R_1 considers α_1 and R_2 considers α_2 as the causal weight of $Z \rightarrow X_1$ (what is now β_2) such that $\alpha_1 \gg \alpha_2$. The same reasoning applies as with causal perception when constructing the interventional distribution $P(Y|X_1, X_2, do(Z := z))$. The difference is that both receivers arrive at (14) after intervening Z with but each has a distinct $\beta_4 \cdot \beta_2 \cdot z$ term. This term is $\beta_4 \cdot \alpha_1 \cdot z$ for R_1 and $\beta_4 \cdot \alpha_2 \cdot z$ for R_2, where $\beta_4 \cdot \alpha_1 \cdot z > \beta_4 \cdot \alpha_2 \cdot z$. Such a difference once again can determine which candidate each officer chooses based on Z.

The term in question in both kinds of causal perception represents a potentially lower or higher SAT score because

of the applicants zip code (read, neighborhood composition). We do this on purpose as the link between wealth (captured by Z) and aptitude (captured by X_1) is a devise one and of interest (Kleinberg et al. 2019).

The central idea with the two kinds of causal perception is that they account for why the two receivers in Definition 4.1 disagree on the poset of implied distributions from a SCM $\mathcal{M}$. The disagreement, by definition, comes from each receiver's SCM: with Definition 4.2 the disagreement stems from different graphical structures while with Definition 4.3 from different model parametrizations. These competing SCMs represent the individual experience of each receiver. Here we take for granted the construction and elicitation of these individual experiences. Notably, though, while in Definition 3.1 we made use of the representativeness heuristic to better understand individual experience in a probabilistic way, there is a lack for an equivalent framework to support Definition 4.1. Therefore, in the next section we propose a causal modeling framework that motivates a receiver's individual experience as a SCM $\mathcal{M}$.

5 Toward a Causal Modeling Framework

How can we describe the process through which a receiver $R \in \mathcal{R}$ constructs its SCM $\mathcal{M}_R$? We treat this section as a first approach given the complexity of the task.

5.1 Key Functionalities

Two processes are central to this framework: categorization and signification. While *categorization* entails sorting instances or classes into categories, *signification* entails representing the social meanings of the categories describing the instances or classes of interest (Loury 2019). For our purposes, given $X_i, X_j \in \mathbf{X}$ categorization implies listing the items that describe X_i and X_j while signification implies specifying how these items relate X_i and X_j. Notably, in signifying any two variables in $\mathbf{X}$ we want to specify causal-like statements between them.

To illustrate these processes, for instance, if X_1 represents SAT scores, then categorizing X_1 could amount to listing items such as "standardized testing", "tutoring", and "performance under pressure". Similarly, signifying X_1 relative to zip code Z could amount to stating that "household income" and "school district" are positively related to "standardized testing". The items "household income" and "school district" are obtained from signifying Z.

Definition 5.1 *(Categorization) Let $\Theta^R(X = x)$ denote the categorization of $X = x$ by receiver R such that:*

$$\Theta^R(X = x) = \{\theta_1^X, \dots, \theta_n^X\}$$

where θ_i^X is the i-th item or category that receiver R associates to X. It is possible for $\Theta^R(X = x) = \emptyset$.

We define the categorization set as:

$$\vartheta^R(\mathbf{X} = \mathbf{x}) = \left\{\Theta^R(X_i = x_i)\right\}_{i=1}^{p} \tag{16}$$

where $p = |\mathbf{X}|$. Each $R \in \mathcal{R}$ comes with its own categorization set. It is implied that X is a variable in the SCM $\mathcal{M}$ used by R. For readability, we write $\Theta^R(X)$ and ϑ^R.

Definition 5.2 *(Signification) Let $\Phi^R(X_i, X_j)$ denote the signification of X_i, X_j by receiver R such that:*

$$\Phi^R(X_i, X_j) = \hat{\Phi}^R(\Theta^R(X_i), \Theta^R(X_j))$$

is one of $X_i \overset{\beta}{\to} X_j$ or $X_j \overset{\beta}{\to} X_i$, which reads as "$X_i$ causes X_j with effect β" or vice versa. It is possible for $\Phi^R(X_i, X_j) = \emptyset$. The idea is that receiver R evaluates the items describing the i-th and j-th variables, derived from its categorization set, and from (the intersection of) these items derives (or not, if the result is $\emptyset$) a causal statement between them that depends on the kind of causal perception.

Under structural causal perception (Definition 4.2) the $\Phi^R(X_i, X_j)$ is causal statement on the cause-effect pair and the causal effect. Under parametrical causal perception (Definition 4.3) the $\Phi^R(X_i, X_j)$ is causal statement only on the causal effect. We define the signification set as:

$$\varphi^R = \left\{\left\{\Phi^R(X_i, X_j)\right\}_{j\neq i, j=1}^{p}\right\}_{i=1}^{p} \tag{17}$$

where $p = |\mathbf{X}|$. Each $R \in \mathcal{R}$ comes with its own signification set that is based on its own categorization set. It is implied that X_i, X_j are variables in the SCM $\mathcal{M}$ used by R.

Implementation-wise, Definition 5.1 relates to knowledge representation formalisms, such as ontologies (Gruber 1995) or knowledge graphs (Hogan et al. 2022) where X is the class or concept and $\Theta^R(X)$ its attributes or properties. Categorization would then amount to building a semantic model (for an ontology) or a network of facts (for a knowledge graph) that R has for $\mathbf{X}$. Similarly, Definition 5.2 relates to methods that argue over represented knowledge, such as logic argumentation (Besnard, Cordier, and Moinard 2014) or relational learning (Salimi et al. 2020), among other techniques for argumenting over knowledge. Signification would then amount to arguing how two variables relate causally to each other based on the affinity between the items describing them. Under structural causal perception the answer would be structural, and under parametrical causal perception the answer would be parametric.

These two definitions are first approaches to capture how the receiver R derives its SCM $\mathcal{M}_R$. Together, ϑ^R and φ^R motivate the construction of $\mathcal{M}_R$. The general idea here is that these two high-level processes are to Definition 4.1 what the representativeness heuristic is to Definition 3.1.

5.2 Illustrative Examples

Let us showcase how two receivers might construct their corresponding SCMs through the processes of categorization and signification, leading to causal perception.

Example 5 *(Categorization) In Example 1, scenario (ii), let us assume that the admissions officer R_1 breaks the tie between the two applicants with the same suitability score G from the ML model f using the applicants' SAT scores X_1, high-school GPA X_2, and address Z. Recall from Example 4 that it is implied that these applicants have the same X_1 and*

X_2. *We define* ϑ^{R_1} *of applicants'* X_1, X_2, *and* Z *as:*

$$\begin{aligned}\vartheta^{R_1} =&\{\Theta^{R_1}(X_1), \Theta^{R_1}(X_2), \Theta^{R_1}(Z)\}\\ =&\{\{\textit{tutoring, expensive, performative}\},\\ &\{\textit{discipline, school funding, potential}\},\\ &\{\textit{family income, school district}\}\}.\end{aligned}$$

The first line states the variables R_1 *categorizes and the second line the descriptors for each variable according to* R_1. *Additionally, for later use in the upcoming examples we define a second admissions officer* R_2 *that has a similar categorization set to* R_1 *but with* $\Theta^{R_2}(X_1) = \emptyset$, *which means that:* $\vartheta^{R_2} = \{\Theta^{R_1}(X_2), \Theta^{R_1}(Z)\}$.

Example 6 *(Signification for Structural Perception) Based on* ϑ^{R_1} *and* ϑ^{R_2} *from Example 5, we define the corresponding signification sets below. Assume due to the college's bylaws that an admissions officer determines* Y *using* X_1 *and* X_2 *only. Hence,* $X_1 \to Y$ *and* $X_2 \to Y$ *are given and* $Z \to Y$ *is not allowed; these cause-effect pairs are provided to and shared by* R_1 *and* R_2. *For* R_1*:*

$$\begin{aligned}\varphi^{R_1} =&\{\Phi^{R_1}(Z, X_1), \Phi^{R_1}(Z, X_2), \Phi^{R_1}(Z, Y),\\ &\Phi^{R_1}(X_1, X_2), \Phi^{R_1}(X_1, Y), \Phi^{R_1}(X_2, Y)\}\\ =&\{\{\hat{\Phi}^{R_1}(\{\textit{family income, school district}\},\\ &\{\textit{tutoring, expensive, performative}\}),\\ &\hat{\Phi}^{R_1}(\{\textit{family income, school district}\},\\ &\{\textit{discipline, school funding, potential}\}), \emptyset,\\ &\hat{\Phi}^{R_1}(\{\textit{tutoring, expensive, performative}\},\\ &\{\textit{discipline, school funding, potential}\}\}),\\ &X_1 \to Y, X_2 \to Y\}\\ =&\{Z \to X_1, Z \to X_2, X_2 \to X_1, X_1 \to Y, X_2 \to Y\}\end{aligned}$$

where the first equality states the pair of variables R_1 *signifies; the second the items of each variable within each pair; and the third the cause-effect ordering for each pair based on these descriptors. For readability, we omit the* β*'s superscripts in the final line. For* R_2 *we have a similar signification set* φ^{R_2} *with the exception of* $\Phi^{R_2}(Z, X_1) = \emptyset$ *as* $\Theta^{R_2}(X_1) = \emptyset$. *This results in the distinct DAGs discussed in Example 4 for structural causal perception.*

Example 7 *(Signification for Parametrical Perception) Similar to Example 6, consider scenario (ii) from Example 1. Suppose* R_1 *and* R_2 *now agree on* $\mathcal{G}_{R_1}$. *We formalize the scenario where two admissions officers consider the potential of the same applicant differently. Assume both officers have similar signification sets* φ^{R_1} *and* φ^{R_2}, *except for* $Z \to X_1$*:*

$$\begin{aligned}\Phi^{R_1}(Z, X_1) &= \hat{\Phi}^{R_1}(\{\textit{family income, school district}\},\\ &\qquad \{\textit{tutoring, expensive, performative}\})\\ &= Z \xrightarrow{\beta_1} X_1\\ \Phi^{R_2}(Z, X_1) &= \hat{\Phi}^{R_1}(\{\textit{family income, school district}\},\\ &\qquad \{\textit{expensive, performative}\})\\ &= Z \xrightarrow{\beta_2} X_1\end{aligned}$$

such that $\beta_1 > \beta_2$. *This results in the distinct causal weights discussed in Example 4 for parametric causal perception.*

In all three examples the address Z acts as a proxy for socioeconomic background. This is intentional as our analysis is based on the original example from Kleinberg et al. (2019) in which an admissions officer decides between two applicants with the same SAT score but from different neighborhoods. We kept the "controversy" from the original example as it shows how a variable like Z could condition an officer's decision-making based on how she reasons about Z conditional on her individual experience.

5.3 Additional Related Work

We extend Section 1.2 by discussing causal perception within the fair causal ML literature.

Abstractions, Categorization, and Signification. Given a SCM $\mathcal{M}$ and its DAG $\mathcal{G}$ for $P(\mathbf{X})$, it is possible to imagine movements between low (or micro) and high (or macro) levels of causal abstraction. Works on causal abstraction (Rubenstein et al. 2017; Beckers, Eberhardt, and Halpern 2019; Massidda et al. 2023) study how causal reasoning materializes and is preserved between levels. These works focus mainly on the notion of consistency, studying whether the how we reason on one level is consistent with the how we reason on another level given a SCM $\mathcal{M}$.[3] For example, suppose that we have one SCM describing the flow of water particles in a river and another SCM describing the flow of the river itself: consistency implies that intervening the water particles toward a specific direction should be equivalent to intervening the river toward that same direction.

With the processes of categorization (Definition 5.1) and signification (Definition 5.2), we point at the concept of causal abstraction by allowing the receivers to go from the random variable X (a high-level representation) to its n descriptors $\{\theta_1^X, \cdots, \theta_n^X\}$ (a low-level representation) and relate these to another variable's descriptors. We do so by assuming a basic structure for X and without formalizing the causal reasoning that takes place in $\mathcal{M}$ and $\mathcal{G}$ across these two levels of abstraction. Our focus is conceptual and centered on knowledge representation, but we share the goal of formalizing how an agent's causal reasoning is impacted by the possible abstractions comprised in a SCM.

Hu and Kohler-Hausmann (2020) question whether a SCM can capture the meaning behind a sensitive random variable, such as gender, by formalizing it as a single node in a DAG. Their work does not discuss causal abstraction, which should account for this critique, but notably argues for a molecular structure to sensitive attributes, such that there is gender (the high-level representation) and more granular nodes related to it (the low-level representation) in the DAG that simultaneously define gender as concept. Both categorization and signification are inspired by Hu and Kohler-Hausmann (2020)'s discussion. Mossé et al. (2025) provide a similar conceptual discussion based solely on causal abstractions and focused on discrimination testing. We are unaware of other fair causal ML works that discuss different levels of causal reasoning. We add to this line of work by

[3] See Rubenstein et al. (2017, Definition. 3; Theorem. 6) and Beckers, Eberhardt, and Halpern (2019, Definition 3.1) for details.

considering how multiple agents can disagree on the abstraction levels when looking at the same (sensitive) attribute and how that leads to perception.

Counterfactual Fairness and Colliding Worlds. Given a SCM $\mathcal{M}$ and its DAG $\mathcal{G}$, counterfactual fairness (CF) (Kusner et al. 2017) establishes that the observed outcome should be the same as the counterfactual outcome when intervening the sensitive attribute. It remains the leading causal fairness metric (Makhlouf, Zhioua, and Palamidessi 2020). By definition, these two outcomes belong, respectively, to the observational and counterfactual distributions that are both included in $\mathcal{P}_{\mathbf{X}}$ (4). Extensions to CF, such as path-specific CF (Chiappa 2019), also deal with Equation (4) as these always compare the observed distribution with respect to some interventional distribution given $\mathcal{M}$ and $\mathcal{G}$.

CF is with respect to a single SCM. In a companion paper to Kusner et al. (2017), Russell et al. (2017) consider the robustness of CF when multiple world views, as in multiple SCMs describing $P(\mathbf{X})$, are used to compute CF. Although multiple SCMs can be problematic as different SCMs lead to different CF results for the same $P(\mathbf{X})$ (Binkyte et al. 2022), Russell et al. (2017) and later works (see, e.g., Kilbertus et al. (2019) on hidden confounders) study the robustness of CF with a focus on the single SCM: all robustness claims are relative to a SCM of interest. The goal is to avoid disagreement by "colliding" the worldviews into one.

With causal perception we consider the opposite scenario from Russell et al. (2017). We move away from colliding worlds by expecting disagreement among the receivers. Method-wise, we could indeed measure the robustness of a fairness metric like CF under causal perception (Definition 4.1) given these previous works. However, we diverge from them conceptually by considering the setting in which disagreement is allowed or even encouraged. To the best of our knowledge, what would amount to CF robustness in our setting remains unexplored. Our causal perception definition is our contribution to this new line of work on alternative CF robustness claims. A similar argument applies to cases in which we rely on a SCM $\mathcal{M}$ and DAG $\mathcal{G}$ but may allow multiple perspectives from stakeholders, such as discrimination testing (Álvarez and Ruggieri 2023, 2025) and synthetic biased data generation (Baumann et al. 2023).

6 Perception and Fairness

Finally, we explore three fair ML areas that benefit from the proposed formalization of (causal) perception.

6.1 Situated Bias

In previous sections, we avoided the term "bias" despite the connection between perception as a cognitive phenomenon and biased human decision-making (Kahneman 2011). This choice was motivated by two considerations. First, the notion of bias often associated to our chosen view of perception (recall, Section 1.2) does not align directly with the definition commonly used in algorithmic fairness. While the former refers to a departure from rational decision-making as described by basic probability laws (Kahneman 2011), the latter refers to "demographic disparities in algorithmic systems that are objectionable for societal reason" (Barocas, Hardt, and Narayanan 2023). The latter interpretation is usually the primary concern in fairness-related discourse. Second, building on this distinction, the notion of bias in algorithmic fairness often carries a negative connotation as one aims to capture objectionable disparities. There is nothing, in principle, negative being captured by Definitions 3.1 and 4.1. Both definitions simply formalize a disagreement in probabilistic and causal probabilistic reasoning between two individuals. We must, thus, be more precise by "situating" the bias (Haraway 1988).

The disagreement between receivers R_i and R_j may lead to bias, but to speak of the kind of bias that drives algorithmic unfairness we must not only have a disagreement but also a preference on which interpretation of the information $\mathbf{X}$ is preferred. Similar to how Haraway (1988) argues that all objective knowledge is based on a partial view of the problem,[4] our framework can be used to situate bias by defining one receiver's (causal) probabilistic interpretation of $\mathbf{X}$ as the reference interpretation. The underlying idea is that since individual experience, which we construct through a SCM $\mathcal{M}$, shapes decision-making, then we can use our formalization to explore settings in which we cannot reach an agreement on what is fair or want to be explicit on what fairness looks like within a specific context.

Example 8 *In Examples 6 and 7 we ca define R_1 as the representative fair receiver given its reasoning behind X_1 and Z. R_1 would then represent the desired decision-maker for breaking potential ties between applicants in terms of the suitability score G of the ML model f.*

6.2 Loaded Attributes

Similarly, in previous sections we did not discuss sensitive attributes, like gender or race, and their role in perception. We argue that these attributes are prone to induce perception, as they are summaries of complex historical and social processes (Bonilla-Silva 1997; Sen and Wasow 2016). As such, they are likely to affect how individuals are perceived by others. We are referring, for example, to the conceptual difference between describing an individual as female versus feminine (Hu and Kohler-Hausmann 2020), both of which are based on the attribute gender. Female refers to a category of gender while feminine refers to a set of behavioral expectations attributed to females: i.e., the biological or phenotypic label versus the socially constructed identity. We describe these attributes as loaded because they are likely to lead to different interpretations among receivers. If $X \in \mathbf{X}$ is a loaded attribute, then it should be easier for a receiver R to evoke individual experience about X.

We regard loaded attributes as attributes that thrive on stereotypes of social categories shared and maintained by the receivers. A social category is the result of classifying people into groups over shared perceived identities (Bowker and Star 1999). We refer to a social category as a social construct when the classification is also used purposely

[4] Objectivism is a function of what we choose to see (situation), how we choose to see it (location), and from where we choose to see it (position), which are telling of privilege and power relations.

to enforce exclusionary policies (Mallon 2007). Sensitive attributes are examples of social constructs. A stereotype refers to the cognitive representation people develop about a particular social category, based around beliefs and expectations about probable behaviors, features and traits, which translate into implicit or explicit attitudes that materialize into bias (Beukeboom and Burgers 2019; Johnson 2020).

The main difficulty, as discussed in Section 5.3, is how to model causally such distinction and, further, how it translates into a potential disagreement among the decision makers. The causal perception framework offers a way to formalize these overloaded attributes as elements of individual experience from the perspective of the receiver. The SCMs condition how receivers interpret information, which is particularly important given the potential biases they may hold toward certain social groups.

Example 9 *As implied in Examples 6 and 7, the admissions officers' consideration of the zip code, Z, of each applicant serves as an attempt to infer their socioeconomic background. By associating Z with X_1 and X_2 and drawing on stereotypes (like correlations between wealth and SAT scores), an officer may develop biased inferences, which ultimately lead to favor one applicant over the other.*

6.3 The Human-in-the-Loop

The next example, while centered on automated decision-making (ADM), illustrates the same interpretative challenges discussed in the previous ones for scenarios *(i)* and *(ii)*. We view the admissions officer as one receiver and her supervisor as another receiver. A natural question to ask then is "What happens if they disagree on the k applicants selected by f?" in scenario *(iii)*. Such situation highlights a broader issue with ML pipelines: the final decision often culminates in human oversight. From a fairness perspective, it is essential to develop frameworks that accommodate subjective interpretations and context-aware notions of fairness (Srivastava, Heidari, and Krause 2019; Yaghini, Krause, and Heidari 2021; Jung et al. 2021).

Example 10 *Consider scenario (iii) in Example 1. Given the top-k applicants by the ML model f, suppose that the admissions officer is asked by the college to explain for the k chosen applicants by f. The officer uses xAI tools, like feature importance methods, to understand the model's outcome and explain it to her supervisor. Assume that the supervisor also has access to f and the same xAI tools. Suppose the officer and her supervisor interpret differently the feature importance explanations.*

Example 10 is based on studies such as Bertrand et al. (2022) that consider the limitations of xAI methods under the threat of human cognitive biases. This divergence between officer and supervisor highlights a key challenge in xAI: how can we design explanations that minimize interpretive ambiguity among users? Similar to nudging in behavioral economics (Gigerenzer 2018), which was inspired by Tversky and Kahneman's work on cognitive biases (Thaler and Sunstein 2008), future xAI tools should aim to anticipate a range of plausible user interpretations and proactively guide users toward a shared, intended understanding. The causal perception framework allows to formalize such set of user-based interpretations in a way that is compatible with ML applications and xAI tools.

7 Conclusion

In this work, we treated perception as a disagreement between individual probabilistic interpretations of the same bit of information, and defined perception of probabilities and perception of implied probabilities. For the latter, which we termed causal perception, we used structural causal models (SCMs) to formalize disagreement between observed and interventional (read, hypothetical) distributions. Further, we formalized two kinds of causal perception based on when individuals disagree on the causal graph (structural perception) and when they disagree on the causal effect(s) given the same casual graph (parametrical perception). Furthermore, we proposed a first approach to how individuals construct these disagreeing individual SCMs through the processes of categorization and signification.

We also outlined fairness research areas that benefit from the proposed framework. Notably, causal perception is useful in contexts in which multiple interpretations and representations of information occur by experts. It is, in turn, useful for problems involving diverse stakeholders that interact with the ML system. The framework allows to formally position fairness in such a setting as well as to reformulate sensitive attributes as loaded with perceptions.

This paper is, above all, conceptual. Next steps include, for instance, using ontologies (Gruber 1995) and relational learning (Salimi et al. 2020)—with Kahneman and Miller (1986)'s norm theory in mind—to construct in practice the SCMs that denote individual experience. Revisiting causal perception under Pearl's three levels of reasoning (here, we simplified it to two levels) is also a next step (Bareinboim et al. 2022). In general, future work should validate the framework empirically through case studies. Additionally, causal perception assumes an explicit level of causal reasoning among individuals that may not be realistic as well as assumes SCMs as individual prior knowledge. Future work should evaluate how practical and scalable these individual SCMs are for the context in question.

Ethical Statement

We did not face ethical challenges when drafting this paper.

Acknowledgments

We are indebted to Riccardo Massida for his discussions, insights, and suggestions. We would like to thank also Mayra Russo, Maria-Esther Vidal, and Andrea Pugnana for their feedback in earlier versions of this paper. Salvatore Ruggieri is supported by the European Union (EU)'s Horizon Europe research and innovation program for the project FINDHR (g.a. 101070212). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU. Neither the EU nor the granting authority can be held responsible for them.

References

Agrawal, A.; Gans, J.; and Goldfarb, A. 2019. *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Álvarez, J. M.; Bringas-Colmenarejo, A.; Elobaid, A.; Fabbrizzi, S.; Fahimi, M.; Ferrara, A.; Ghodsi, S.; Mougan, C.; Papageorgiou, I.; Lobo, P. R.; Russo, M.; Scott, K. M.; State, L.; Zhao, X.; and Ruggieri, S. 2024. Policy advice and best practices on bias and fairness in AI. *Ethics Inf. Technol.*, 26(2): 31.

Álvarez, J. M.; and Ruggieri, S. 2023. Counterfactual Situation Testing: Uncovering Discrimination under Fairness given the Difference. In *EAAMO*, 2:1–2:11. ACM.

Álvarez, J. M.; and Ruggieri, S. 2025. Counterfactual Situation Testing: From Single to Multidimensional Discrimination. *J. Artif. Intell. Res.*, 82: 2279–2323.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference*, volume 36 of *ACM Books*, 507–556. ACM.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Baumann, J.; Castelnovo, A.; Crupi, R.; Inverardi, N.; and Regoli, D. 2023. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *FAccT*, 1002–1013. ACM.

Beckers, S.; Eberhardt, F.; and Halpern, J. Y. 2019. Approximate Causal Abstractions. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, 606–615. AUAI Press.

Bengio, Y. 2019. NeurIPS 2019 Posner Lecture: From System 1 Deep Learning to System 2 Deep Learning. https://nips.cc/Conferences/2019/ScheduleMultitrack?event=15488.

Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *AIES*, 78–91. ACM.

Besnard, P.; Cordier, M.; and Moinard, Y. 2014. Arguments Using Ontological and Causal Knowledge. In *FoIKS*, volume 8367 of *Lecture Notes in Computer Science*, 79–96. Springer.

Beukeboom, C.; and Burgers, C. 2019. How stereotypes are shared through language: A review and introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7: 1–37.

Binkyte, R.; Makhlouf, K.; Pinzón, C.; Zhioua, S.; and Palamidessi, C. 2022. Causal Discovery for Fairness. In *AFCP*, volume 214 of *Proceedings of Machine Learning Research*, 7–22. PMLR.

Bonilla-Silva, E. 1997. Rethinking Racism: Toward a structural interpretation. *American Sociological Review*, 465–480.

Booch, G.; Fabiano, F.; Horesh, L.; Kate, K.; Lenchner, J.; Linck, N.; Loreggia, A.; Murugesan, K.; Mattei, N.; Rossi, F.; and Srivastava, B. 2021. Thinking Fast and Slow in AI. In *AAAI*, 15042–15046. AAAI Press.

Bordalo, P.; Coffman, K.; Gennaioli, N.; and Shleifer, A. 2016. Stereotypes. *The Quarterly Journal of Economics*, 131(4): 1753–1794.

Bowker, G.; and Star, S. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Cabitza, F.; Campagner, A.; and Basile, V. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. In *AAAI*, 6860–6868. AAAI Press.

Caraban, A. K.; and Karapanos, E. 2020. The '23 ways to nudge' framework: designing technologies that influence behavior subtly. *Interactions*, 27(5): 54–58.

Chiappa, S. 2019. Path-Specific Counterfactual Fairness. In *AAAI*, 7801–7808. AAAI Press.

Costello, F. J. 2009. How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, 22: 213–234.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under Human Assistance. In *AAAI*, 2611–2620. AAAI Press.

Dittadi, A.; Träuble, F.; Locatello, F.; Wuthrich, M.; Agrawal, V.; Winther, O.; Bauer, S.; and Schölkopf, B. 2021. On the Transfer of Disentangled Representations in Realistic Settings. In *ICLR*. OpenReview.net.

Echterhoff, J. M.; Yarmand, M.; and McAuley, J. J. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *CHI*, 161:1–161:9. ACM.

Fahimi, M.; Russo, M.; Scott, K. M.; Vidal, M.; Berendt, B.; and Kinder-Kurlanda, K. 2024. Articulation Work and Tinkering for Fairness in Machine Learning. *Proc. ACM Hum. Comput. Interact.*, 8(CSCW2): 1–23.

Gigerenzer, G. 2018. The Bias Bias in Behavioral Economics. *Review of Behavioral Economics*, 5(3-4): 303–336.

Goldenberg, I.; and Webb, G. I. 2019. Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowl. Inf. Syst.*, 60(2): 591–615.

Goldstein, E. B. 1996. *Sensation and Perception*. Wadsworth, 4th edition.

Goodfellow, I.; Bengio, Y.; Courville, A.; and Bengio, Y. 2016. *Deep learning*. MIT press Cambridge.

Gregory, R. L. 1970. *The intelligent eye*. McGraw-Hill.

Gruber, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum. Comput. Stud.*, 43: 907–928.

Guidotti, R.; Monreale, A.; Giannotti, F.; Pedreschi, D.; Ruggieri, S.; and Turini, F. 2019. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intell. Syst.*, 34(6): 14–23.

Haraway, D. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3): 575–599.

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; Ngomo, A. N.; Polleres, A.; Rashid, S. M.; Rula, A.; Schmelzeisen, L.; Sequeda, J. F.; Staab, S.;

and Zimmermann, A. 2022. Knowledge Graphs. *ACM Comput. Surv.*, 54(4): 71:1–71:37.

Hu, L.; and Kohler-Hausmann, I. 2020. What's sex got to do with machine learning? In *FAT**, 513. ACM.

Johnson, G. M. 2020. The structure of bias. *Mind*, 129(516): 1193–1236.

Jung, C.; Kearns, M.; Neel, S.; Roth, A.; Stapleton, L.; and Wu, Z. S. 2021. An Algorithmic Framework for Fairness Elicitation. In *FORC*, volume 192 of *LIPIcs*, 2:1–2:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D. 2021. NeurIPS 2021: A Conversation on Human and Machine Intelligence. https://nips.cc/virtual/2021/invited-talk/22284.

Kahneman, D.; and Miller, D. T. 1986. Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2): 136–153.

Kahneman, D.; Rosenfield, A. M.; Gandhi, L.; and Blaser, T. 2016. Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harvard Business Review*.

Kilbertus, N.; Ball, P. J.; Kusner, M. J.; Weller, A.; and Silva, R. 2019. The Sensitivity of Counterfactual Fairness to Unmeasured Confounding. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, 616–626. AUAI Press.

Kleinberg, J. M.; Ludwig, J.; Mullainathan, S.; and Sunstein, C. R. 2019. Discrimination in the Age of Algorithms. *CoRR*, abs/1902.03731.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *NIPS*, 4066–4076.

Loftus, J. R. 2024. Position: The Causal Revolution Needs Scientific Pragmatism. In *ICML*. OpenReview.net.

Loury, G. 2019. Why Does Racial Inequality Persist? Culture, Causation, and Responsibility. *The Manhattan Institute*.

Madras, D.; Pitassi, T.; and Zemel, R. S. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *NeurIPS*, 6150–6160.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. Survey on Causal-based Machine Learning Fairness Notions. *CoRR*, abs/2010.09553.

Mallon, R. 2007. A field guide to social construction. *Philosophy Compass*, 2(1): 93–108.

Massidda, R.; Geiger, A.; Icard, T.; and Bacciu, D. 2023. Causal Abstraction with Soft Interventions. In *CLeaR*, volume 213 of *Proceedings of Machine Learning Research*, 68–87. PMLR.

Mossé, M.; Schechtman, K.; Eberhardt, F.; and Icard, T. 2025. Modeling Discrimination with Causal Abstraction. *arXiv preprint arXiv:2501.08429*.

Mozannar, H.; Lang, H.; Wei, D.; Sattigeri, P.; Das, S.; and Sontag, D. A. 2023. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. In *AISTATS*, volume 206, 10520–10545. PMLR.

Palomba, F.; Pugnana, A.; Álvarez, J. M.; and Ruggieri, S. 2025. A Causal Framework for Evaluating Deferring Systems. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, 2143–2151. PMLR.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Elsevier.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A.; and Tomsett, R. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW1): 83:1–83:22.

Rong, Y.; Leemann, T.; Nguyen, T.; Fiedler, L.; Qian, P.; Unhelkar, V. V.; Seidel, T.; Kasneci, G.; and Kasneci, E. 2024. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(4): 2104–2122.

Rubenstein, P. K.; Weichwald, S.; Bongers, S.; Mooij, J. M.; Janzing, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2017. Causal Consistency of Structural Equation Models. In *UAI*. AUAI Press.

Ruggieri, S.; Álvarez, J. M.; Pugnana, A.; State, L.; and Turini, F. 2023. Can We Trust Fair-AI? In *AAAI*, 15421–15430. AAAI Press.

Ruggieri, S.; and Pugnana, A. 2025. Things Machine Learning Models Know That They Don't Know. In *AAAI*, 28684–28693. AAAI Press.

Russell, C.; Kusner, M. J.; Loftus, J. R.; and Silva, R. 2017. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *NIPS*, 6414–6423.

Salimi, B.; Parikh, H.; Kayali, M.; Getoor, L.; Roy, S.; and Suciu, D. 2020. Causal Relational Learning. In *SIGMOD Conference*, 241–256. ACM.

Scantamburlo, T.; Baumann, J.; and Heitz, C. 2024. On prediction-modelers and decision-makers: why fairness requires more than a fair prediction model. *AI & SOCIETY*, 1–17.

Schölkopf, B. 2022. Causality for Machine Learning. In *Probabilistic and Causal Inference*, volume 36 of *ACM Books*, 765–804. ACM.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards Causal Representation Learning. *CoRR*, abs/2102.11107.

Sen, M.; and Wasow, O. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19(1): 499–522.

Sobel, J. 2020. Signaling Games. *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, 251–268.

Spence, M. 1973. Job Market Signaling. *The Quarterly Journal of Economics*, 87(3): 355–374.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *KDD*, 2459–2468. ACM.

Tentori, K. 2022. What Can the Conjunction Fallacy Tell Us about Human Reasoning? In *Human-Like Machine Intelligence*, 449–464. Oxford University Press.

Thaler, R.; and Sunstein, C. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Tversky, A.; and Kahneman, D. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.

Tversky, A.; and Kahneman, D. 1981. Judgments of and by Representativeness. Technical Report 3, Defense Technical Information Center.

Tversky, A.; and Kahneman, D. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4): 293.

van Steenkiste, S.; Locatello, F.; Schmidhuber, J.; and Bachem, O. 2019. Are Disentangled Representations Helpful for Abstract Visual Reasoning? In *NeurIPS*, 14222–14235.

Woodward, J. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yaghini, M.; Krause, A.; and Heidari, H. 2021. A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness. In *AIES*, 1023–1033. ACM.

Yang, S. C.-H.; Folke, T.; and Shafto, P. 2025. The Inner Loop of Collective Human–Machine Intelligence. *Topics in Cognitive Science*, 17: 248—267.