International Journal of Approximate Reasoning 190 (2026) 109608

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Data speak but sometimes lie: A game-theoretic approach to data bias and algorithmic fairness

Chiara Manganini *, Esther Anna Corsi , Giuseppe Primiero

University of Milan, Via Festa del Perdono, 7, Milan, 20122, Italy

ARTICLE INFO

Keywords:
Machine learning
Bias
Data quality
Data-driven inference
Non-monotonic logic

ABSTRACT

In the present work, we develop a novel information-theoretic and logic-based approach to data bias in Machine Learning predictions and show its relevance in the specific context of fairness evaluation. We frame predictions made on biased data as Ulam games, which formalise key aspects of data-driven inference, and from which a variation of the rational non-monotonic consequence relation can be defined. We investigate this framework to model how differential levels of noise in input features impact Machine Learning predictions. To the best of our knowledge, this is the first game-theoretic formalisation of ML unfairness.

1. Introduction

Over the past decade, numerous frameworks have been proposed to reason about the unfairness of Machine Learning (ML) systems. Some of these accounts have stressed that ML unfairness can be caused by a biased representation of certain sensitive groups within the data [2,3]. Unfair predictions are often described as the consequence of "*mismeasured* proxies [of the sensitive attribute]", or emerging when "attributes […] *misrepresent* the true behaviour of the users" [1, pp. 4, 7],[1] where concepts like *mis*-measurement and *mis*-representation clearly pertain to the vocabulary of data quality, in particular concerning the dimension of data correctness. We hence argue that there is value in framing the problem of unfair predictions from this data-quality perspective, in terms of the relationship between *incorrect* ML input data and *unfair* ML outputs.

Similarly to [2,3], in what follows we aim to focus on a specific type of noise that affects how individuals are represented as inputs of an ML algorithm. More specifically, for such noise to affect the fairness of the predictions, it must possess two fundamental properties. First, it must be *unevenly distributed* between sensitive groups (such as those individuated on the basis of gender, race, and age), as only errors that are non-uniform between sensitive groups can impact their outcomes differentially [4]. Second, this noise needs to be *systematic*, in the sense that it must consistently overestimate or underestimate the true values of an ML input feature and therefore cannot, unlike random noise, be averaged out by merely adding more records. Conversely, we assume that the corresponding ML labels do not contain any data bias [2,3].

Hence, the question is whether there is a viable way to model, measure, and account for the contribution of data bias to ML predictions. Here, taking seriously Friedler *et al.*'s suggestion that it might be "possible to take a more information-theoretic perspective on the nature of transformations between [construct and observed] spaces" [2, p. 5], we precisely explore an information-theoretic and logic-based approach to data bias. Specifically, we exploit the setting of a two-player game, called the Ulam game [5], in which

* Corresponding author.
E-mail address: chiara.manganini@unimi.it (C. Manganini).

[1] Emphasis added.

https://doi.org/10.1016/j.ijar.2025.109608
Received 28 March 2025; Received in revised form 24 November 2025; Accepted 29 November 2025
Available online 6 December 2025

one player selects a secret element from a finite search space, and the other tries to discover it by asking a series of yes-no questions. Both players know that *at most* $0 \leq m < \infty$ answers can be untruthful. For example:

> Someone thinks of a number between one and one million (which is just less than 2^{20}). Another person is allowed to ask up to twenty questions, to each of which the first person is supposed to answer only yes or no. Obviously the number can be guessed by asking first: Is the number in the first half million? Then again reduce the reservoir of numbers in the next question by one-half, and so on. Finally the number is obtained in less than $log_2(1,000,000)$. Now suppose one were allowed to lie once or twice, then how many questions would one need to get the right answer? [5, p. 281]

As Pelc pointed out in [6], the two-party communication that occurs in the Ulam game can be interpreted as communication between a sender (Responder) and a receiver (Questioner), who communicate via a noisy channel in which the original message sent by the sender becomes distorted. In this setting, the maximum number of lies Responder can perform corresponds to the limitation on errors (noise) present in the message received by the receiver.

In previous work [7], an alternative interpretation of the Ulam game has been proposed. Specifically, the game was used to model data-driven reasoning in science. Scientists, like Questioner, aim at discovering which scientific hypothesis H^* holds (the secret) in a search space $\mathcal{H}$ that can be understood as a set of mutually exclusive and exhaustive scientific hypotheses. They pursue this objective by performing experiments, the outcomes of which can be interpreted as Nature's answers to yes-no questions (i.e., experiments). Thus, Nature plays the role of Responder, and the untruthful answers Nature gives can be seen as errors inevitably slipping into experimental data. One fundamental aspect of data-driven scientific reasoning captured by the Ulam game is its *gradedness*, meaning that the process of hypothesis rejection comes in degrees: a hypothesis is said to be *temporarily rejected* whenever a single (possibly untruthful) question-and-answer entails its negation, while it is said to be *rejected* if it has been temporarily rejected at least $m + 1$ times.

In the present work, we adapt the Ulam game to yet another data-driven scenario, namely that of ML predictions. A prediction can be understood, in fact, as the outcome of the game, where Questioner corresponds to a trained ML model, and Responder to a new input. Questioner/ML Model wants to know the true label H^* of a data instance by asking a sequence of yes-no questions, each of which corresponds to a binary ML feature.[2] At each new question-and-answer, Questioner updates the hypotheses' degrees of rejection accordingly. This description connects most naturally with an important class of ML algorithms: decision trees. Each internal node poses a yes-no question about an input feature, and each root-to-leaf path encodes a structured interrogation strategy leading Questioner to the secret hypothesis.

In this framework, data errors can be interpreted as the untruthful responses given by Responder/Nature. More specifically, the type of data bias we described earlier can be modelled using a variant of the classic version of the game, where Responder can choose to deliver her answers through different channels, each characterised by a different noise level. In a scenario with no data bias, Responder consistently uses the same channel to provide the responses, regardless of the sensitive group to which the data point belongs. These answers can be noisy, but this noise is uniformly distributed between individuals of different sensitive groups – in Ulam terms, each data point "is playing the same game". In contrast, in a situation of data bias, the malicious Responder chooses to use a differentially noisy (i.e., more or less noisy) channel to deliver information about the individuals that belong to a specific sensitive group.

To illustrate this model of data bias, we use a real-world medical scenario in which a racially biased data-generating mechanism produces significantly different observations for factually very similar patients, who subsequently end up receiving significantly different medical treatments. In line with broader efforts to apply non-monotonic reasoning to ML settings (see, e.g., [8]), we use a variation of the rational non-monotonic consequence relation inspired by the Ulam game to quantify this disparity in a novel way, without assuming any causal knowledge of the data-generating process and relying solely on the input data and the model's associated decisions.

The paper is organised as follows. After introducing a working example from the medical domain (Section 2), we introduce some preliminary notions about the Ulam game (Section 3) and show how this game-theoretic formalisation provides a suitable model for ML decision-making (Section 4). In Section 5, we first define a rule of constrained monotonicity for the Ulam game, and then give its generalised version for weighted evidence. After connecting our proposal with already existing related work (Section 6), in Section 7 we introduce a proof-theoretic system that captures key aspects of the Ulam game and subsequently establish a completeness result for the Ulam consequence relation. Finally, we conclude by discussing the general applicability of the proposed framework for reasoning about ML unfairness (Section 8).

2. A medical working example

To illustrate our methodology, we start with an example taken from a report on the equity of medical devices. The document was released by UK's National Health Service (NHS) [9]. In the report, a fictitious story is used to illustrate the serious consequences of racially biased medical measurements within an ML-aided medical scenario. In the example, a pulse oximeter – an optical device that sends light waves of various frequencies through the patient's skin to make measurements of underlying physiology – systematically overestimates the true oxygen saturation level in individuals with higher levels of melanin in the skin. The narrative is realistic, as there is extensive literature on the clinical impact of mismeasured saturation levels in black patients [10,11].

[2] Or to a multi-valued feature that has been adequately binarised, e.g., by using a one-hot encoding process.

The story is about two patients of different skin colours who are screened by an automatic risk-assessment system, that we assume to be an ML model, to decide which one needs intensive care (IC henceforth) the most. Since the example is set during the COVID-19 pandemic, we can easily assume that there are very limited resources in the hospital, so that *only one of the two patients* can be admitted to the intensive care unit at the time of initial evaluation. Crucially, the measured level of blood oxygen in the dark-skinned patient (not recognised as biased by the doctors) causes her to obtain delayed care.

> Steven, a 58-year-old **White** British man, has been experiencing worsening **breathlessness, fever and chest pain**. He arrives at his office, sweaty and breathless, and with a new cough. His manager calls 999. Paramedics attend and find him so breathless that they rush him to the nearest emergency department. A PCR test for COVID-19 is positive, and his chest x-ray shows severe **pneumonitis**. A pulse pulse oximeter shows his **blood oxygen levels are low**. The emergency team assesses him using a standard clinical risk prediction score, which shows a **very high score**. The hospital's protocol for scores this high means that he is referred to the intensive care unit […] to receive **respiratory support** […]. He is seriously ill for a few days, but makes a good recovery. He goes home after 10 days.
> Yvonne, a 60-year-old **Black** British woman of Caribbean heritage, has **similar symptoms** and is barely able to manage her job as a cleaner because of her **breathlessness**. As an agency worker, she has no sick pay, so she delays calling 999 until she is home from work. Paramedics find her so breathless that they take her straight to the nearest emergency department. A PCR test for **COVID-19 flags positive**, and her chest x-ray shows **severe pneumonitis**. A pulse pulse oximeter shows her **blood oxygen levels are just within normal** range. The emergency team assesses her using a clinical risk prediction score, which comes back **high, but not sufficiently high to trigger a referral to intensive care**. **Further tests** show her **oxygen levels are worse** than the pulse pulse oximeter suggests. She is finally **moved to intensive care** […]. Her oxygen levels are now critically low, so she is put on a **ventilator** to save her life. [9, p. 39][3]

It is important to notice that, from a medical standpoint, a low level of oxygen in the blood is associated with a high risk of requiring intensive care. Therefore, data *do speak in favour* of Steven over Yvonne in this story. Based on the two patients' data, the clinician who is unaware of the measurement error produced by the pulse oximeter will presumably deem it *fair* to choose Steven over Yvonne, as suggested by the "clinical risk prediction score" in the quote, that we here assume to be produced by an ML model.[4] The decision process seems to meet two important procedural aspects of fair decisions:

Principle 1 (Justifiability of Evaluation)**.** *The features used to evaluate individuals have been selected based on domain knowledge and their use is justified by their having proven causal influence on the individuals' outcome.*

Principle 2 (Consistency of Evaluation)**.** *A consistent method of evaluation has been used for both individuals, as the same information has been collected about them.*

According to the clinician who assesses the ML outcome, choosing Steven over Yvonne is fair in virtue of the justifiable and consistent procedure used to decide such an outcome. In what follows, we want to examine whether this procedure is actually so. While we will assume the Principle 1 to hold – at least defeasibly, based on the clinician's current medical knowledge – we will focus on Principle 2, to the conclusion that it is ultimately *not satisfied*. It will be shown in fact, that not the same information, but merely the same observations, have been collected about the two patients. Our proposed information-theoretic approach shows that the two patients' data are associated with different levels of information. To quantify it, we start by introducing the key logical aspects of the Ulam game.

3. Preliminaries

Starting from the idea that data can reject hypotheses, in [7], *degree of rejection* functions have been introduced. A generic degree of rejection is defined over the multiset[5] of data $\mathcal{M}_D$ times the set of hypotheses and has values in the positive real numbers or ∞. The use of multiset is necessary to accommodate the fact that seeing the same data once is not the same as seeing it multiple times. Suppose that $\mathcal{H} = \{H_1, \dots, H_n\}$ and $\mathcal{D} = \{d_1, \dots, d_l\}$ are two sets of propositional variables standing for *hypotheses* and *data*, respectively. Then, $\mathrm{Fm}_{\mathcal{H}}$ and $\mathrm{Fm}_{\mathcal{D}}$ are the sets of propositional formulas generated by closing $\mathcal{H}$ and $\mathcal{D}$, respectively, under the usual connectives $\wedge, \vee, \neg$. Note that the definition of degree of rejection does not tell us how to compute it, but only which properties a function should satisfy to be a degree of rejection.

[3] Emphasis added.

[4] We are assuming that the algorithm is not "aware" that even a normal oxygen level is associated with high risk for a specific group of patients (dark-skinned ones). This might happen for a number of reasons. The system might have been trained on unbiased measurement methods of blood oxygen (like through blood tests, for instance), or the training set did not contain enough records of dark-skinned patients, or again, the ML model was specifically designed to incorporate medical-specific knowledge about the importance of blood oxygen level to evaluate the risk of serious respiratory issues in patients. Furthermore, we are also assuming that the ML model is not "aware" of the race of the two patients, as this is not an input feature.

[5] A multiset is a generalisation of a set that allows multiple occurrences of the same element. Unlike sets, where each element appears only once, multisets keep track of the number of times each element appears (its multiplicity).

Table 1
An intuitive formalisation of the medical observations collected for assessing the risk of the two patients of the example illustrated in Section 2. In the table, we have assumed that both patients' age (58 and 60, respectively) is below a certain fixed risk threshold, arbitrarily set to 70 for the sake of the example.

#	Questions	Steven's Answers	Yvonne's Answers
0	70+ years old?	No	No
1	Breathlessness?	Yes	Yes
2	Fever?	Yes	Yes
3	Chest pain?	Yes	Yes
4	Cough?	Yes	Yes
5	COVID-19 test?	Yes	Yes
6	Severe pneumonitis?	Yes	Yes
7	Low blood oxygen?	Yes	No
	ML Decision	Move to IC	Not move to IC
8	Follow-up Test 1?	-	Yes
9	Follow-up Test 2?	-	Yes
10	Follow-up Test 3?	-	Yes

Definition 1. We say that $r: \mathcal{M}_D \times \mathcal{H} \to \mathbb{R}_+ \cup \{\infty\}$ is a *degree of rejection* if, for any $\gamma, \delta \in \mathrm{Fm}_D$, $\Delta \in \mathcal{M}_D$ and $H \in \mathcal{H}$, the following hold:

1. If $T, \gamma \models \delta$, then $r_\gamma(H) \geq r_\delta(H)$.
2. If $T, H \models \delta$ then $r_\delta(H) = 0$.
3. If $T, \delta \models \neg H$, then $r_\delta(H) > 0$.
4. If $r_\Delta(H) \neq \infty$, then $r_\Delta(H) = \sum_{\delta \in \Delta} r_\delta(H)$.

The classical formula T, defined as

$$T := \bigvee_{H \in \mathcal{H}} H \ \wedge \bigwedge_{H, H' \in \mathcal{H},\ H \neq H'} H \to \neg H',$$

formalizes the assumption that the set of hypotheses $\mathcal{H}$ is exhaustive and mutually exclusive.

Once the degree of rejection has been defined, we can identify the set $\hat{\mathcal{H}}_\Delta$ of hypotheses in $\mathcal{H}$ which are *least rejected* by the data in Δ:

$$\hat{\mathcal{H}}_\Delta = \mathrm{argmin}_{\mathrm{H} \in \mathcal{H}^{\mathrm{f}}_\Delta} \mathrm{r}_\Delta(\mathrm{H}). \tag{1}$$

where $\mathcal{H}^f_\Delta$ is the set of hypotheses with a finite degree of rejection. We have then recalled all the necessary elements for defining the *RJ-consequence relation* (where *RJ* stands for *rejection*).

Definition 2 (RJ-consequence)**.** Let Δ be a multiset of formulas in Fm_D and $\varphi \in \mathrm{Fm}_\mathcal{H}$. We say that φ is an *RJ-consequence of* Δ, written $\Delta \mathrel{|\!\sim_{RJ}} \varphi$, if $T, H \models \varphi$, for each $H \in \hat{\mathcal{H}}_\Delta$.

Based on Definition 2, in [7] a family of non-monotonic consequence relations is introduced. One of them, in particular, finds its motivation in the play process of the Ulam game. This consequence relation is defined by instantiating the degree of rejection that was defined in Definition 1 as follows. Suppose that H^* stands for the secret and m for the maximum number of lies Nature can tell. Then, the Ulam rejection degree r^u is defined as:

$$r^u_\delta(H) = \begin{cases} \infty & \text{if } T \models \neg\delta, \\ 1 & \text{if } T, H \models \neg\delta, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$r^u_\Delta(H_i) = \begin{cases} \infty & \text{if } H_i \in \mathcal{H} \text{ and } \sum_{\delta \in \Delta} r^u_\delta(H^*) > m \\ \infty & \text{if } H_i \neq H^* \text{ and } \sum_{\delta \in \Delta} r^u_\delta(H_i) > m \\ \sum_{\delta \in \Delta} r^u_\delta(H_i) & \text{otherwise.} \end{cases} \tag{3}$$

Thus, the rejection degree has value ∞ only if we go against the game's rules (the secret H^* has been rejected more than m times) or the sum of the rejection degrees of the single $\delta \in \Delta$ is strictly greater than m.

4. Modelling ML decisions as Ulam games

4.1. Modelling ML features

Our language is composed of two (finite) sets of propositional variables $D = \{d_1, \ldots, d_l\}$ and $\mathcal{H} = \{H_1, \ldots, H_n\}$, respectively corresponding to the feature space (or data) and to the target label space (or hypotheses).

In the oximeter example, for instance, the set of hypotheses is $\mathcal{H} = \{IC, \neg IC\}$, which corresponds to the possible prediction of the ML model. Moreover, we denote with $Q_i(Yv)$ the i-th question of Table 1 whenever it is referred to Yvonne. Similarly, $Q_i(St)$ denotes the i-th question whenever it is referred to Steven. Then, $Q_i(Yv) =$ "Yes" indicates that the answer to question i on Yvonne is positive. To enhance readability for positive answers to question i we will write $Q_i(a)$, and $\neg Q_i(a)$ for negative answers, where a denotes a generic patient. Thus, $\mathcal{D} = \{Q_i(a) \mid i = 0, \dots, 10\}$. Accordingly, since the input features used by the ML model to assess the patients' respiratory risk correspond to questions 0–7 in Table 1, they can be rephrased as follows:

$$\begin{aligned}\Delta_{Yv} &= \{\neg Q_0(Yv), Q_1(Yv), Q_2(Yv), Q_3(Yv), Q_4(Yv), Q_5(Yv),\\ &\quad Q_6(Yv), \neg Q_7(Yv)\}\\ \Delta_{St} &= \{\neg Q_0(St), Q_1(St), Q_2(St), Q_3(St), Q_4(St), Q_5(St),\\ &\quad Q_6(St), Q_7(St)\}\end{aligned}$$

Since the level of admissible noise in the data Δ_{Yv} and Δ_{St} is not known, we initially assume it to be greater than the number of questions. Otherwise, Δ_{Yv} and Δ_{St} would yield very similar (or even the same) rejection degrees for both hypotheses. If r is the degree of rejection function defined by the functions (2) and (3), then:

$$r_{\Delta_{Yv}}(IC) = 2, \text{ and } r_{\Delta_{Yv}}(\neg IC) = 6,$$

$$r_{\Delta_{St}}(IC) = 1, \text{ and } r_{\Delta_{St}}(\neg IC) = 7.$$

According to a deterministic function that maps the data onto hypotheses, it is possible to rewrite the elements in Δ_{Yv} and Δ_{St} in the language of $\mathcal{H}$. In this example, we assume that for all $i = 0, \dots, 7$

$$Q_i(a) \equiv IC$$

while

$$\neg Q_i(a) \equiv \neg IC.$$

Thus, from every positive answer to questions 0–7 of Table 1 it classically follows that the generic patient a should go to intensive care (IC). The contrary follows from negative answers. Consequently:

$$\Delta_{Yv} \equiv IC^6 \wedge \neg IC^2$$

$$\Delta_{St} \equiv IC^7 \wedge \neg IC$$

where IC^k stands for $\bigwedge_{i=1,\dots,k} IC_i$ and $\neg IC^k$ stands for $\bigwedge_{i=1,\dots,k} \neg IC_i$. The pedics of IC_i and $\neg IC_i$ have the sole role of counting the occurrences of IC and $\neg IC$.

In ML terms, these equivalences convey the information one can obtain through feature importance techniques such as SHAP [12]. In fact, $Q_i(a) \equiv IC$ expresses that observing that a satisfies predicate Q_i (e.g., has fever) "contributes" to hypothesis IC rather than $\neg IC$. However, it is worth noting that the analogy needs to be refined, as here every observation is assumed to have the same importance, which is unrealistic for the large majority of ML scenarios. This assumption will be dropped in Section 5.2. In addition, if $m = 0$ and the search space consists of only two hypotheses, then we would need only one question-and-answer to determine whether a patient should be admitted to the IC. In practice, each of the questions $Q_0, \dots, Q_7$ in isolation is not sufficient for the practitioner to make a decision.

4.2. Modelling ML decisions

By ML decision-making, we mean a decision process that involves the use of an ML algorithm to define a function $\hat{f} : \mathcal{X} \to [0, 1]$ which associates individuals with a probabilistic prediction (such as a risk score). This prediction is used to prioritise individuals, given a decision function d and a fixed threshold τ:

$$d(\vec{x}) = \begin{cases} 1 & \hat{f}(\vec{x}) \geq \tau \\ 0 & \text{otherwise.} \end{cases}$$

We model this ML-enabled decision-making as an Ulam game in the following way. First, each ML prediction corresponds to a play of the game where Questioner corresponds to the trained ML model, and Responder to the data point $\vec{x} = \{x_1, \dots, x_n\}$ it is fed with. Questioner wants to find the data point's true label y between mutually exclusive hypotheses (in what follows these will be assumed to be just two, H and $\neg H$, to model standard risk-assessment scenarios). Questioner asks a number of yes-no questions about $\vec{x}$, each corresponding to a binarised feature x_i. After every answer, Questioner updates the degree of rejection of H and $\neg H$ accordingly. Table 2 summarises the mapping between the fundamental concepts of Ulam game and ML predictions.

In our simplified example with only two patients, a decision needs to be made on which patient, between Yvonne and Steven, should be admitted to the IC unit. It is important to note that both patients show symptoms of a respiratory condition. In fact, IC is the least rejected hypothesis for both, i.e., $\hat{\mathcal{H}}_{\Delta_{Yv}} = \hat{\mathcal{H}}_{\Delta_{St}} = \{IC\}$. Therefore, by Definition 2, $\Delta_{Yv} \vdash_{RJ} IC$ and $\Delta_{St} \vdash_{RJ} IC$. Nonetheless, we also know that $d(Steven) = 1$ and $d(Yvonne) = 0$, where 1 corresponds to "Move to IC" and 0 to its negation. In fact,

Table 2
Correspondence between the Ulam Game and ML predictions.

Ulam game	ML interpretation
Questioner	ML system
Questions	Feature space $\mathcal{X}$
Responder	Nature
Answers	Data point $\vec{x}$
Different channels	Data bias
Lies	Incorrect data
Secret	True label y

we assumed in Section 2 that *only one* patient can be moved to the IC unit, due to limited resources. Denoting by a a generic patient in $\{Yvonne, Steven\}$ and by $-a$ the other one, d is defined as follows:

$$d(a) = \begin{cases} 1 & r_{\Delta_a}(IC) < r_{\Delta_{-a}}(IC) \\ 0 & r_{\Delta_a}(IC) > r_{\Delta_{-a}}(IC) \\ \texttt{undec} & \text{otherwise} \end{cases} \tag{4}$$

Had Steven's and Yvonne's rejection degrees been the same, the decision function would have been undecidable ($d(a) = \texttt{undec}$). In fact, the available data Δ_a would have been insufficient to justifiably move either patient to intensive care.

4.3. Modelling data bias

If we assume the oximeter measurements to be accurate (i.e., error $\varepsilon = 0$), from the data we have, Yvonne's clinical picture differs from Steven's. In reality, we know this is not the case, as the two patients have in fact similar conditions. In fact, the device has captured Steven's (factually low) blood oxygen level in a sufficiently accurate way, while it has overestimated Yvonne's (factually low, as well). More specifically, her blood oxygen level is masked as normal (meaning, above a given threshold t assumed to be the same for both patients). We do not know how much this bias ε amounts to, but we know that had Yvonne's blood oxygen level been lower than $t - \varepsilon$, she would have been flagged as $Q_7(Yv)$. Since this has not been the case, Yvonne's true saturation level is between $(t - \varepsilon)$ and t.

To model data bias, we modify the classic version of the Ulam game by introducing the notion of *channels* into the game. Similarly to previous applications of the Ulam game to error-tolerant communication [6], we interpret each channel as corresponding to a different number of lies available to Responder/Nature in the classic version of the Ulam game [13]. We can think of each channel as corresponding to a different noise level in the communication *medium* used by Responder/Nature to answer Questioner. It might be that the answer Questioner receives through a channel is totally accurate (no lies), it might be that Responder/Nature can use up to m lies, or it might even be that Responder/Nature's lies are related to the outcome of an aleatory experiment – in this latter case, we are talking of a channel with probabilistic lies [14,15].

Crucially, every data point is associated with its specific channel or set of channels (ideally, one for each feature and independent of each other). In the oximeter example, we know that the device underestimates the level of oxygen in dark-skinned patients, while it works sufficiently well for light-skinned ones. In medical terms, a pulse oximeter is put on Yvonne's finger, and then the data appearing on the screen is transcribed as "low" or "normal" (depending on the value of threshold t) on her medical file and given as input to the ML algorithm. In Ulam terms, this corresponds to asking Responder Question 7. The number of lies actually used by Responder within Steven's game is $m_{St} = 0$, while for Yvonne $m_{Yv} > 0$ (note that we are able to reconstruct this only because we know that $\neg Q_7(Y)$ should have been $Q_7(Y)$). For instance, if $m_{Yv} \leq 2$ – i.e., at most 2 of the observations in Δ_{Yv} are incorrect – then the compatible correct data would be:

1. $\Delta'_{Yv} \equiv IC^6 \wedge \neg IC^2$ (no lie has been said);
2. $\Delta''_{Yv} \equiv IC^8$ (2 lies on $\neg IC$);
3. $\Delta'''_{Yv} \equiv IC^4 \wedge \neg IC^4$ (2 lies on IC);
4. $\Delta''''_{Yv} \equiv IC^5 \wedge \neg IC^3$ (1 lie on IC, 1 lie on $\neg IC$).
5. $\Delta'''''_{Yv} \equiv IC^7 \wedge \neg IC$ (1 lie on $\neg IC$);
6. $\Delta''''''_{Yv} \equiv IC^5 \wedge \neg IC^3$ (1 lie on IC);

Thus, if we apply the decision function (4) to cases (1–6), we would obtain:

1. $d(\Delta'_{Yv}) = 0$;
2. $d(\Delta''_{Yv}) = 1$;
3. $d(\Delta'''_{Yv}) = 0$;
4. $d(\Delta''''_{Yv}) = 0$.
5. $d(\Delta'''''_{Yv}) = \texttt{undec}$
6. $d(\Delta''''''_{Yv}) = 0$

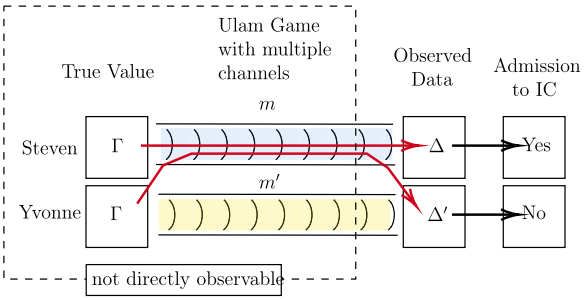

Fig. 1. Initial and Unfair Position. Observations Δ and Δ' have been produced through two channels characterised by different levels of noise, respectively m and m', with $m' > m$. Despite this, as the red arrows indicate, the unaware Questioner misassociates observations Δ' with the wrong channel, assuming them to be more truthful than what they actually are. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

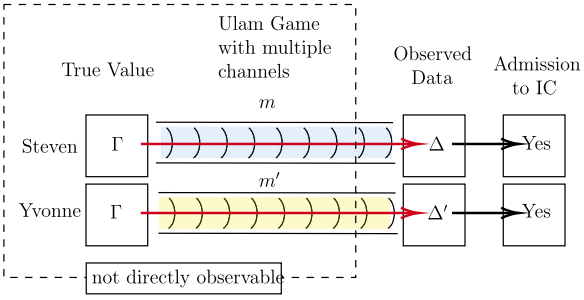

Fig. 2. Final and Fair Position. Observations Δ and Δ' have been produced through two channels characterised by different levels of noise, respectively m and m', with $m' > m$. The red arrows indicate that Questioner is correctly associating Δ and Δ' to their true noise level. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

When $m_{Y_v} \leq 2$, it can no longer be inferred that Steven should be chosen over Yvonne for intensive care. In fact, supposing that Responder can use up to 2 lies, it is no longer true that $d(\Delta_{Y_v}) = 0$ for all correct evidence ${\Delta^*}_{Y_v}$ compatible with Δ_{Y_v}.[6] Specifically, the observations at hand would be compatible with the situation in which Yvonne is actually *more in need of intensive care* than Steven – this happens when both lies are on $\neg IC$, corresponding to case (2.). In other words, it might be the case that her oxygen level is below the threshold t and $\neg Q_7(Y)$ as result of a lie, just as much it could be the case that her oxygen level is factually above the threshold t and $\neg Q_7(Y)$ is accurate. In a situation like this, data bias corresponds to the presence of more than one channel connecting Question 7 to the patients. Given these channels, unfairness corresponds to the situation in which a data point is misassociated with the wrong channels, while fairness corresponds to the situation in which this misassociation does not occur. Figs. 1 and 2 illustrate how the medical example could produce these unfair and fair decisions, respectively.

Let us now see how a proof-theoretic investigation of the consequence relation capturing the Ulam rejection degree can hint at a way to test a generic ML algorithm for this misassociation.

5. Model-based ML testing within Ulam

In the classic version of the Ulam game, we can detect a lie whenever we receive contradictory information from Nature. For example, if we ask the same question twice and obtain different answers, then one of the two must be false, and we can infer that Nature has lied. Suppose l lies have been spotted, then the number of lies m is *at least* l ($m \geq l$).

In the version of the game we consider, however, we assume that m is fixed but it is not common knowledge, i.e., only Nature knows it. This makes it difficult to recognise incorrect data: any data we collect may or may not be so, as we cannot know when the lies are over. Nonetheless, we know that *there must be* a point in the game at which Responder runs out of lies, since m is finite.

In a game setting like this, although we cannot recover m from the data obtained by questioning Nature, it is nevertheless possible to define an order over the hypotheses based on how many times they have been rejected. Moreover, since the number of lies available to Nature will eventually end, we know that from a certain point onwards (that we do not know), Nature's answers will all be correct. This means that the more questions we ask, the more robust – i.e., invariant to new and possibly erroneous information – our rejection function becomes. Let us suppose, for instance, that we could gather all possible information about Yvonne and use it as input of an ML model accordingly trained. Similarly to the oximeter reading, some other observations would be noisy. Yet, the usual assumption of independence prevents them to be *all noisy in the same direction and with the same skew*. While we would not know m and, therefore,

[6] Here, $*\in \{', '', ''', '''', ''''', '''''' \}$.

Table 3
Let us denote with d the cardinality of our data Δ. We compute the number of checks needed to verify the validity of $(\text{UMO})^i$ with $i \in \{1, 2, 3\}$ for a generic d and for some specific d.

$(\text{UMO})^i$	Number of checks for a generic cardinality d of Δ	$d = 10$	$d = 50$	$d = 100$
$i = 1$	d	10	50	100
$i = 2$	$\binom{d}{2} = \frac{d(d-1)}{2}$	45	1125	4950
$i = 3$	$\binom{d}{3} = \frac{d(d-1)(d-2)}{6}$	120	19600	161700

we could not be able to establish at which point Responder has run out of lies, we would still notice that, as observations are added, one hypothesis starts being inferred more and more consistently from the data.

The broader our set of observations, the more robust our rejection order on hypotheses becomes and only an increasingly exceptional set of observations could change it. Hence, measuring the distance in rejection degrees between the least rejected hypothesis and the second leats rejected one can indicate whether the observations are in fact converging on a hypothesis. Even though assuming that errors in the data might occur at maximum only a fixed and finite number of times is an oversimplification on the usual distribution of errors in data, in the following, we show how this robustness is effectively captured by the rule of Ulam constrained MOnotonicity (UMO) first (Subsection 5.1), and then by its Generalised version (gUMO) for weighted evidence (Subsection 5.2). In Section 8, we discuss in more detail how variants of the Ulam game – such as versions incorporating both communication channels and probabilistic half-lies – could provide more sophisticated models of the bias present in the data.

5.1. The (UMO) rule of constrained monotonicity

The uRJ-consequence relation obtained by instantiating r in Definition 2 with r^u satisfies the following rule of constrained monotonicity.

$$\frac{\Delta \vdash \psi \quad \{\Delta \setminus \{\delta\} \vdash \psi\}_{\delta \in \Delta}}{\Delta, \varphi \vdash \psi} \text{ (UMO)}$$

Let us now see a possible way to read (UMO). Suppose that whatever there is on the left-hand side of the consequence relation $\vdash$ stands for the data we have, while on the right-hand side of $\vdash$, we find what follows from the data Δ. Therefore, suppose that ψ follows from the data in Δ, in symbols, $\Delta \vdash \psi$. Since Δ formalizes the data we are considering, it is defined as a multiset of formulas, and each $\delta \in \Delta$ can be seen as an observation. Suppose now to remove one single observation from Δ and verify whether from $\Delta \setminus \delta$ we can still infer ψ. If $\{\Delta \setminus \{\delta\} \vdash \psi\}_{\delta \in \Delta}$, then (UMO) tells us that we can add any other observation φ to Δ, but the data in Δ are strong enough to still infer ψ.

This rule of constrained monotonicity can be generalised concerning the cardinality of the set of observations that we remove from Δ. In (UMO), we remove *one* observation at the time from Δ, and then we add only *one* observation. Let d be the cardinality of Δ, then for every set of indexes $I \subseteq [d]$, we can identify those of cardinality 2 and generalize (UMO) as follows.

$$\frac{\Delta \vdash \psi \quad \{\Delta \setminus \{\bigcup_{i \in I} \delta_i\} \vdash \psi\}_{|I|=2, I \subseteq [d]}}{\Delta, \bigcup_{\substack{j \in J \\ |J| \leq 2}} \varphi_j \vdash \psi} \text{ (UMO)}^2$$

Note that the indexes in the conclusion of the rule only have the role of counting the maximum number of new evidence that we can add to the premise Δ and still inferring ψ. Generalising for every $n \leq d$, we can define the following rule.

$$\frac{\Delta \vdash \psi \quad \{\Delta \setminus \{\bigcup_{i \in I} \delta_i\} \vdash \psi\}_{|I|=n, I \subseteq [d]}}{\Delta, \bigcup_{\substack{j \in J \\ |J| \leq n}} \varphi_j \vdash \psi} \text{ (UMO)}^n$$

The rule (UMO) is then recovered imposing $n = 1$, and whenever $n = 0$, the two premises and the conclusion are identical. For $n = d$, $(\text{UMO})^n$ holds only if ψ is a tautology. The evidence $\{\varphi_1, \ldots, \varphi_n\}$ added to Δ could be interpreted as an addition of (potentially disrupted) data that Δ is able to tolerate. The rule $(\text{UMO})^n$ can be used to test a generic ML algorithm. Suppose that n is the maximum n for which $(\text{UMO})^n$ is valid; then the ML prediction's robustness should be proportional to it. As n grows, the more robust the ML prediction becomes.

In Table 3, we exemplify some results on the number of checks one needs to perform for some increasing levels of the $(\text{UMO})^i$ rule. In Fig. 3, an illustration of the UMO is given.

5.2. A generalised (UMO) rule for weighted evidence

In the setting analyzed in [7], data are formalised through multisets to reflect the fact that seeing the same data once is not the same as seeing it multiple times. However, for computing the rejection degree, the relevance of each observation for the prediction is not taken into account. One simple way to generalize the rejection degree is to multiply every $r_\delta(H)$ by a weight w_δ whose value is in [0, 1], and that indicates how relevant δ is to the prediction.

Thus, the *weighted rejection degree* can be defined as follows.

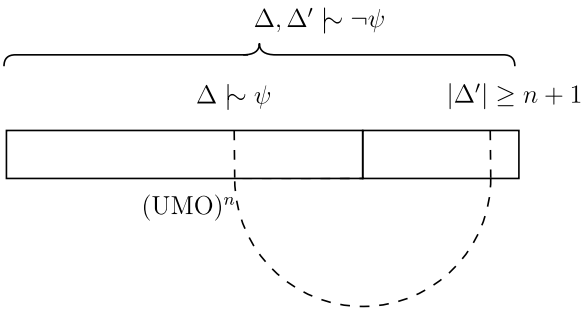

Fig. 3. In the example, Δ infers ψ satisfying $(\text{UMO})^n$. If we add Δ the set of observations Δ', with $|\Delta'| \geq n+1$ and $\Delta' \mid\sim_{RJ} \neg\psi$, then the initial conclusion gets overtuned..

Definition 3. Let r be a *rejection degree function* and $w: \mathcal{D} \times \mathcal{H} \to [0,1]$ a weighted function that reflects the importance of each observation δ relative to a hypothesis H. Then, the *weighed rejection degree function* $r^w: \mathcal{M}_{\mathcal{D}} \times \mathcal{H} \to \mathbb{R}_+ \cup \{\infty\}$ is defined as

$$r^w_\Delta = \sum_{\delta \in \Delta} r_\delta(H) w_\delta(H)$$

If we consider a weighted rejection degree function instead of a rejection degree function, then Definition 2 can be easily extended. The (UMO) rule formalizes a scenario in which the rejection level of the least rejected hypotheses is sufficiently low compared to the rejection level of all other hypotheses. This ensures that adding a single extra piece of information does not alter the order of rejection of the hypotheses and, ultimately, the set of least rejected hypotheses and what follows from them.

In the context of the Ulam game, since one extra piece of information can change the level of rejection of at most one, the second premise of the rule is equivalent to requiring that the difference between the rejection degree of the least rejected hypotheses and the rejection degree of the rest of them is at least 2. If this is the case, then we can add any piece of information to a generic Δ, and the level of rejection of any non-least rejected hypotheses will necessarily be greater than the rejection degree of those minimally rejected by Δ. Thus, we have found a condition that guarantees that $\hat{H}_{\Delta,\varphi} \subseteq \hat{H}_\Delta$. However, if we consider a weighted Ulam-based rejection degree function, it might be that $\hat{H}_{\Delta,\varphi} \nsubseteq \hat{H}_\Delta$. Thus, the second premise of (UMO) needs to be refined. For any single evidence δ and $H \in \mathcal{H}$, $r^u_\delta(H) \in \{0,1\}$. Thus, if we consider the weighted extension of r^u, denoted by r^{uw}, we have that $r^{uw}_\delta(H) = r^u_\delta(H) w_\delta(H) \in [0,1]$. In general, for any value of w_δ, we have that $r^u_\Delta(H) \geq r^{uw}_\Delta(H)$ and while $r^u_\Delta(H) \in \mathbb{N} \cup \{\infty\}$, $r^{uw}_\Delta(H) \in \mathbb{R}_+ \cup \{\infty\}$.

Suppose that for every $H \in \hat{H}_\Delta$ $r^{uw}_\Delta(H) > 0$. Then it might be the case that there exists $H' \in \mathcal{H} \setminus \hat{H}_\Delta$ s.t. $r^{uw}_\varphi(H') = 0$ and

$$r^{uw}_{\Delta,\varphi}(H') \leq r^{uw}_{\Delta,\varphi}(H) \text{ i.e.}$$

$$r^{uw}_{\Delta,\varphi}(H') = r^{uw}_\Delta(H') + r^{uw}_\varphi(H') = r^{uw}_\Delta(H') \leq r^{uw}_\Delta(H) + r^{uw}_\varphi(H).$$

If $r^{uw}_\Delta(H') \leq r^{uw}_\Delta(H) + r^{uw}_\varphi(H)$, then H' might belong to $\hat{H}_{\Delta,\varphi}$ and $\hat{H}_{\Delta,\varphi} \nsubseteq \hat{H}_\Delta$. Thus, to make sure that $\hat{H}_{\Delta,\varphi} \subseteq \hat{H}_\Delta$, we need to require that the difference of the weighted rejection degrees between any hypotheses not minimally rejected by Δ and the weighted rejection degree between any hypotheses minimally rejected by Δ is strictly greater than the weighted rejection degree of the new information φ on any of the minimally rejected hypotheses, i.e. if for every $H \in \hat{H}_\Delta$ and $H' \in \mathcal{H} \setminus \hat{H}_\Delta$ $r^{uw}_\varphi(H) < r^{uw}_\Delta(H') - r^{uw}_\Delta(H)$, then $\hat{H}_{\Delta,\varphi} \subseteq \hat{H}_\Delta$. The rule (UMO) can then be generalised to (gUMO).

$$\frac{\Delta \mid\sim \psi \quad \{r_\varphi(H) < r_\Delta(H') - r_\Delta(H)\}_{H \in \hat{H}_\Delta, H' \in \mathcal{H} \setminus \hat{H}_\Delta}}{\Delta, \varphi \mid\sim \psi} \text{ (gUMO)}$$

Proposition 1. *(gUMO) is valid for the RJ consequence relation obtained by considering the weighted Ulam-based rejection degree function.*

Proof. Suppose that $\Delta \mid\sim_{RJ} \psi$. Thus, for any $H \in \hat{H}_\Delta$ and $H' \in \mathcal{H} \setminus \hat{H}_\Delta$,

$$r^{uw}_\Delta(H) < r^{uw}_\Delta(H').$$

If for $\varphi \in \mathcal{D}$,

$$r^{uw}_\varphi(H) < r^{uw}_\Delta(H') - r^{uw}_\Delta(H),$$

then

$$r^{uw}_{\Delta,\varphi}(H) = r^{uw}_\Delta(H) + r^{uw}_\varphi(H) < r^{uw}_\Delta(H') < r^{uw}_{\Delta,\varphi}(H'),$$

which implies that the rejection order is preserved after adding φ. Therefore, the least rejected hypotheses remain the same, ensuring that $\Delta, \varphi \mid\sim_{RJ} \psi$. □

Table 4
Weights of Questions 0–10 of Table 1; each weight is a value in the interval $[0, 1]$ that reflects the importance of the corresponding question for hypothesis IC.

#	Questions	Weight
0.	70 + ?	0.85
1.	Breathlessness?	0.67
2.	Fever?	0.62
3.	Chest pain?	0.73
4.	Cough?	0.81
5.	COVID-19 test?	0.78
6.	Severe pneumonitis?	0.59
7.	Low blood oxygen?	0.94
8.	Follow-up Test 1?	0.87
9.	Follow-up Test 2?	0.79
10.	Follow-up Test 3?	0.30

Table 5
Degrees of rejection, gUMO level, minimally rejected hypothesis, and final decision computed on the sequence of questions-and-answers Δ_i for Steven and Yvonne, where $\Delta_i = \{Q_0, \dots, Q_i\}$. Finally, for the sake of the example, let us assume that a prediction is considered acceptable when it is associated with a gUMO level greater than 4.20. Therefore, no additional observations are collected for Yvonne beyond Δ_{10}.

	Steven					Yvonne				
	$r(\neg IC)$	$r(IC)$	gUMO	$\hat{\mathcal{H}}$	d	$r(\neg IC)$	$r(IC)$	gUMO	$\hat{\mathcal{H}}$	d
Δ_0	0	0.85	0.85	$\neg IC$	-	0	0.85	0.85	$\neg IC$	-
Δ_1	0.67	0.85	0.18	$\neg IC$	-	0.67	0.85	0.18	$\neg IC$	-
Δ_2	1.29	0.85	0.44	IC	-	1.29	0.85	0.44	IC	-
Δ_3	2.02	0.85	1.17	IC	-	2.02	0.85	1.17	IC	-
Δ_4	2.83	0.85	1.98	IC	-	2.83	0.85	1.98	IC	-
Δ_5	3.61	0.85	2.76	IC	-	3.61	0.85	2.76	IC	-
Δ_6	4.2	0.85	3.35	IC	-	4.2	0.85	3.35	IC	-
Δ_7	5.14	0.85	4.29	IC	1	4.2	1.79	2.41	IC	0
Δ_8	-	-	-	-	1	5.07	1.79	3.28	IC	0
Δ_9	-	-	-	-	1	5.86	1.79	4.08	IC	0
Δ_{10}	-	-	-	-	0	6.16	1.79	4.37	IC	1

Note that in (gUMO), in contrast with (UMO), the second premise imposes a constraint on which φ can be added to Δ. Based on the distance between the least and the second least rejected hypothesis by data Δ, we can hence define a measure $(\text{gUMO})^j$ of the level of monotonicity that a given set of data can tolerate:

Definition 4 ($(\text{gUMO})^j$)**.** If $j = \min\{r_\Delta(H') - r_\Delta(H) \mid H \in \hat{\mathcal{H}}_\Delta, H' \in \mathcal{H} \setminus \hat{\mathcal{H}}_\Delta\}$, then we say that Δ satisfies $(\text{gUMO})^j$.

Conceptually, $(\text{gUMO})^j$ quantifies the robustness of inference $\Delta \mid\!\sim_{RJ} \psi$ to the addition of new (and possibly corrupted) data. Translated in Ulam game terms, the rule tells us the level of tolerance of the data at hand to possible new lies from Responder. Returning to the pulse oximeter example once again, let us assume that the clinician knows, from previous knowledge, how relevant the questions in Table 1 are for the ML prediction. Therefore, assuming that these weights are transcribed in Table 4, it is possible to reconstruct that Steven's data satisfy a higher level of constrained monotonicity with respect to Yvonne. Specifically, Table 5 shows that, arrived at Question 7, Steven's data satisfy $(\text{gUMO})^{4.29}$ and Yvonne satisfies $(\text{gUMO})^{2.41}$. Crucially, the difference $1.88 = 4.29 - 2.41$ quantifies how robust Steven's prediction is, in comparison to Yvonne's. Such a difference effectively quantifies the impact of the biased oximeter reading on Yvonne's prediction. To obtain a robustness level similar to Steven's, she needs to undergo three more tests for IC (questions 8–10), which may require time, resources, and effort from her part, possibly worsening her medical condition in the meantime. This hints at the more general consideration that in order to reach parity of gUMO levels between sensitive groups in the presence of data bias, more information must be collected for the affected sensitive group. This strategy has already been discussed in relation to some applications of active learning for fair predictions [3,16,17], with which our research shows several connections.

Furthermore, decision (4) was originally defined as a function of $r(IC)$. However, we now suggest a possible improvement by redefining it as a function of gUMO, instead:

$$d(a) = \begin{cases} 1 & \text{gUMO}_a > \text{gUMO}_{-a} \\ 0 & \text{gUMO}_a < \text{gUMO}_{-a} \\ \texttt{undec} & \text{otherwise} \end{cases} \tag{5}$$

This better captures the intuition that the decision of admitting Steven over Yvonne to IC could be overturned as we gather more information about her condition, stopping only when her prediction reaches a sufficient level of robustness – to be established

beforehand, based on the sample. We argue that, only then, Principle 2 is eventually satisfied, making the corresponding decisions fair.

6. Related work

6.1. Ulam game for data-driven reasoning

The Ulam game, also known as the "Rényi-Ulam game" has been independently introduced by Alfréd Rényi [18] and by Stanisław Ulam [5, p. 281]. Whenever $m = 0$, – i.e., whenever the answer of Responder are always truthful – the Ulam game is a sound and complete semantics for classical logic, while whenever $m > 0$, the game is a sound and complete semantics for the $m + 2$-valued Łukasiewicz logic [19,20].

To our knowledge, Cicalese [21] is the first to use the Ulam game to model a computational scenario, namely supervised *learning*. Questioner's goal is to determine the secret Boolean function f^* based on a sequence of given input-output pairs and knowing that up to m of these are incorrect. At every new question, Questioner is given a binary input x and has to guess the value of $f^*(x)$. Responder either confirms or refutes it, having the possibility of lying at most m times. Questioner's goal is to identify the target function f^* with the minimum number of wrong predictions. As the game goes on, Questioner proceeds by progressively pruning away those functions that are not compatible with the available noisy observations (input-output pairs) anymore, within a given noise tolerance threshold m. Our work is in a sense complementary to Cicalese's. In fact, while he uses Ulam to model constrained supervised learning in a noisy environment (i.e., the training phase of an ML algorithm), we use the same game to model how an ML algorithm that has already been trained predicts new inputs. Moreover, while Cicalese focuses on noisy *labels*, we focus on noisy *features*, instead.

Similarly to our work, in [22] the Ulam game is used to model a typical data quality problem faced by crowdsourcing platforms, where humans are employed to annotate images (e.g., labelling a dog's picture based on its breed). Since human labellers can fail their tasks for many reasons such as fatigue, ignorance, and difficulty of the task at hand, the challenge here is to implement an error-tolerant annotation process. To do so, the authors break down the original task into many smaller subtasks (i.e., yes-no questions about the dog's hair colour, length, etc.). Given that each subtask is subject to an independent error, they exploit the Ulam game to find an optimal strategy to distribute these subtasks within a group of human workers in order to obtain reliable labels. A computationally efficient heuristic sampling strategy is proposed and evaluated.

6.2. ML fairness

Our model highlights that a similarity-based characterisation of ML fairness – similar individuals should receive similar outcomes [23] – is deeply problematic in the presence of data bias. Input features are in fact often measured with errors that reflect structural inequalities and track individuals' sensitive attributes. Hence, similar individuals belonging to different sensitive groups might produce different observations depending on the channel used by Responder and, conversely, similar observations might correspond to factually dissimilar individuals belonging to different sensitive groups.

Broadly speaking, two distinct attitudes stand in opposition to the similarity-based conception of ML fairness.

On the one hand, a strategy is that of assuming that *any* difference whatsoever between the sensitive groups in the observed target feature is entirely imputable to a biased data-generating mechanism [2]. While in the Construct Space (i.e., the unobservable space underlying our measurement) features are distributed uniformly across sensitive groups, the data-generating process produces differentially noisier representations of individuals in the Observed Space. Recalling the terminology used by [3], such a bias causes uneven levels of *separability* within sensitive groups, as positive and negative instances become more difficult to distinguish. With this additional assumption, non-discrimination can be achieved at the level of ML predictions in terms of parity of positive outcome rates between the sensitive groups. This parity is measured by fairness metrics such as demographic parity and disparate impact [24].

On the other hand, counterfactual fairness requires that we spell out *in full detail* the biased data generating process, i.e., the way sensitive features (e.g., the patient's race) influence the values of non-sensitive ones (e.g. her saturation level) – along with how they both influence the ML prediction [25]. Only once this is causally transparent, unfairness can be quantified as the difference in outcomes between each data point and its counterfactual with respect to the sensitive attribute. In other words, a fair ML model is one that effectively compensates for the bias in the data, always at the level of the ML predictions. However, one great shortcoming of counterfactual fairness is that the causal knowledge required is largely unattainable in many practical ML scenarios.

6.3. Fairness and ML correctability

Our proposal shows meaningful connections with other works on the correctability of ML predictions. Venkatasubramania and Alfano [26] distinguish between two types of ML correction scenarios: algorithmic recourse and algorithmic appeal.

The former pertains to the situation in which an individual who has received an *unfavourable* ML prediction undertakes strategic actions to obtain a favourable decision from the same ML decision process in the future. Here, recourse corresponds to the effort associated with such a sequence of actions. The notion of recourse has gained great attention in the ML fairness debate, especially in relation to the development of techniques to ensure its parity between sensitive groups [27–29]. However, one general objection that can be moved against adopting equality of algorithmic recourse as a fairness criterion is that the notion of recourse in itself puts the emphasis on the *wrong* aspect of an ML prediction, namely the *possibility* of reverting it in case it is unfavourable. The fundamental question of whether such a prediction is *actually* justified in the first place remains largely untouched by the algorithmic recourse

perspective. By putting much of the stress on the sequence of actions that the single individual can put in place to reverse an adverse prediction, the responsibility of the outcome itself is somehow deflected from the AI system to the specific end-user who has to undertake a number of actions to reverse it [30]. But what about an outcome, like the one received by Yvonne in Section 2, that is ultimately unjustified, since it is produced on an incorrect piece of information? In light of the preliminary discussion on data bias given in Section 1, we know that systematic errors affecting the data-generating mechanism can produce a differentially noisy representation of sensitive groups. Therefore, we argue that the conception of ML correctability returned by the notion of algorithmic recourse is ultimately inadequate.

Algorithmic appeal, on the other hand, pertains to situations in which an ML decision is not simply *unfavourable*, but also *unjustified* because of the presence of inaccuracies in the input data used to produce it [26]. In this case, one clearly wants to challenge the ML decision because it was based on *incorrect* information. Unfortunately, the notion of algorithmic appeal has thus far been largely neglected in the literature, with much of the attention on ML correctability directed instead towards algorithmic recourse. Venkatasubramanian and Alfano themselves, who coined the term, seem to have a somewhat oversimplified understanding of it: "someone may be recorded as having defaulted on a loan when in fact they repaid it on time and in full, or they may be recorded as having no credit history when in fact they have taken out and repaid several loans" [26, p. 288]. Although their example concerns a rather idiosyncratic case of annotation error, our previous discussion points to a more compelling context in which algorithmic appeal becomes essential – namely, that of ML predictions made on *systematically biased* input data.

However, as conceptually compelling as it might be, the practical applicability of algorithmic appeal seems limited as long as we do not have direct access to the true values to rectify the incorrect pieces of information. Cerrato *et al.* [31] have introduced the more practically useful concept of "contextual recourse", to denote the situation in which the individual shows that their prediction should be different in virtue of *additional* pieces of information that were not used for the initial prediction. As the authors suggest, "some level of logical reasoning is […] required in contextual recourse" [31, p. 3]. We hold that the logical machinery presented above exactly addresses this point. Through the Ulam game formalisation, in fact, we have quantified this very correctability as a measure of how robust an inference is to possible errors contained in its premises. As we have shown, this inference robustness is effectively captured by the gUMO level calculated using the gUMO rule. To be deemed fair, an ML model predicting in the presence of data bias should attain similar gUMO levels between sensitive groups, as this grants the prediction process to be consistent, as required by Principle 2.

Finally, in previous work [32], we provided an abstract syntax to express the amount of additional information that is sufficient to obtain a correct prediction for a certain data point, from an initial incorrect one (what we called "correction distance" [32]). In the next section, we recall the basic elements of the approach, introducing some minor departures from the original paper.[7] Our last goal is to establish a completeness result for $\vdash_{RJ}$ with respect to this proof system.

7. A completeness result for $\vdash_{RJ}$

7.1. Preliminaries

In the proof theory, we interpret non-monotonicity as a probabilistic derivability relation implementing the failure of the structural rule of Weakening. We start with introducing the following sets of ML predicates, upon which we will then define the Training and the Test Samples:

Definition 5 (ML Predicates)**.** Two types of predicates are defined:

- $\mathbb{P} = \{P_1, \dots, P_n\}$ as the set of predicates corresponding to the ML features;
- $\mathbb{T} = \{T\}$ as the singleton containing the binary target feature.

Definition 6 (Training Sample)**.** A training sample $\mathcal{S}^{train} = \{\mathbb{D}^{train}, \mathbb{P}^{train}\}$ is such that:

- $\mathbb{D}^{train} = \{b_1, b_2, \dots, b_m\}$ is a finite set of elements of a domain denoting the data points of the training sample;
- $\mathbb{P}^{train} = \mathbb{P} \cup \mathbb{T}$ as defined in Definition 5.

Definition 7 (Test Sample)**.** A test sample $\mathcal{S}^{test} = \{\mathbb{D}^{test}, \mathbb{P}^{test}, ww\}$ is such that:

- $\mathbb{D}^{test} = \{a_1, a_2, \dots, a_m\}$ is a finite set of elements of a domain denoting the datapoints of the test sample;
- $\mathbb{P}^{test} = \mathbb{P}$;
- $ww : \mathbb{P}^{test} \mapsto [0, 1]$ is a total weighting function that assigns to each predicate a score representing its relevance to the prediction. These relevance scores are obtained using feature importance techniques.

[7] The formalisation that we will present here departs from the original paper [32] in the following respects:

- we removed the distinction between protected and non-protected predicates in Definitions 5 and 7,
- we changed the weight function ww (see Definition 7),
- the added Definitions 8 and 9,
- we changed the *(base case)* and $(failure)$ inference rules (see Fig. 5).

$$\frac{}{\{Q(a) \mid Q \in \mathbb{P}^{test}\} :: features}\ base\ case$$

$$\frac{\Gamma^{w} :: features \qquad P(a)^{w'} \notin \Gamma \mid P \in \mathbb{P}^{test} \qquad |w + w'| \leq 1}{\{\Gamma, P(a)\}^{W=w+w'} :: features}\ selection$$

Fig. 4. Construction Rules for the Inductive Selection of Features.

Moreover, to ensure that the weights are comparable and sum to one, we additionally define a normalised and signed weight function as follows.

Definition 8 (Normalised Weights)**.** Given a predicate $P \in \mathbb{P}^{\text{test}}$, we define function $w_{norm} : \mathbb{P}^{test} \mapsto [0, 1]$ as:

$$w_{\text{norm}}(P) = \frac{ww(P)}{\sum_{Q \in \mathbb{P}^{\text{test}}} ww(Q)}$$

Definition 9 (Signed Weights)**.** We define function $\tilde{w} : \mathbb{P}^{test} \times \mathbb{D}^{test} \mapsto [-1, 1]$ as follows:

$$\tilde{w}(P_i, a) := \begin{cases} +\, w_{\text{norm}}(P_i) & \text{if } P_i(a), \\ -\, w_{\text{norm}}(P_i) & \text{otherwise.} \end{cases}$$

This can be clarified as follows: for any datapoint, we check the current list of predicates (i.e., ML features); knowing that a predicate holds for a datapoint will weight in the classification at hand for the value of the predicate; knowing that the predicate does not hold, we take the inverse, so that satisfied predicates contribute positively and unsatisfied ones negatively. $P(a)$ reads as "It is the case that individual a has property P" or "Individual a is P". Conversely, $\neg P(a)$ is read as "It is not the case that individual a has property P" or "Individual a is not P". Note that an important assumption here is that the predicates satisfied by a 'speak in favour' of prediction $T(a)$, conversely, predicates that are not satisfied by a 'speak against' prediction $T(a)$. Finally, for convenience, we rewrite $\tilde{w}(P_i, a)$ with the following shortcut when we are referring to a generic a:

$$w_i := \tilde{w}(P_i, a) \tag{6}$$

The language is then defined by predicative expressions closed under negation, conjunction, disjunction and implication.

Definition 10 (Language)**.**

$\phi := P(a)^{w} \mid \neg\phi$
$\Phi := \phi \wedge \phi \mid \phi \vee \phi \mid \phi \rightarrow \phi$
$\psi := T(a)$ such that $T \in \mathbb{T}$ and $a \in \mathbb{D}^{test}$
$\xi := r \in \mathbb{R}$

Some clarifications are important on the definition of $\phi := P(a)^{w}$ where $w = \tilde{w}(P, a)$ as per (6) and $P \in \mathbb{P}^{test}$. Note that we do not assume a generalised version of the Excluded Middle, as it is sensible to consider only properties that can actually be derived (or their negation can) as a target feature from the current list of available predicates for the datapoint under evaluation. Moreover, the weighted inference relation we introduce requires a weaker than classical definition of negation.

Metavariable ψ denotes the result of the classification $T(a)$ for a predicate T that ranges over the class $\mathbb{T}$ of target predicates. Finally, with ξ, we refer to a real value associated to the degree to which the target predicate can be inferred from the current feature set.

7.2. Inference rules

Two simple rules inductively generate the feature list, thereby defining a procedure that selects ML features sequentially, each accompanied by an associated weight (see Fig. 4).

The induction starts from a list containing a single predicate (*base case*). At any step, an atomic predicative formula of weight w is added (*selection*), where the order in which predicates are added may depend from the availability of incoming information, or could be established by design. The overall weight of a feature list is denoted by $W = \sum(w_1, \ldots, w_n)$, as it is the algebraic sum of the weights of the predicates in it. Each selection allows to update the feature list, where a new weight w is added to the previous sum of weights W, always checking the stopping condition that the summation on the overall weight $|W| \leq 1$, which is the total amount of available information about a datapoint. This additive process returns the overall weight of the selected features.

Let us now clarify how, at each step of the induction, the current list of selected features corresponds to the set of premises used by the inferential system to output a probabilistic prediction.

Definition 11 (Classification)**.** A judgement:

$$\Gamma^{test,W} \mid\!\sim^{trained} \psi_{\xi}$$

denotes a classification where:

$$\frac{}{\Gamma^W, \phi \ |\sim \psi_\xi}\, classification$$

$$\frac{\Gamma^W \ |\sim \psi_\xi}{\Gamma^W \ |\sim \neg\psi_{(-\xi)}}\, failure \qquad \frac{\Gamma^W \ |\sim \neg\psi_\xi}{\Gamma^W \ |\sim \psi \to \bot_\xi}\, safety$$

$$\frac{\Delta^W \ |\sim \psi_{1\xi} \qquad \Gamma^{W'} \ |\sim \psi_{2\chi}}{\{\Delta,\Gamma\}^{W+W'} \ |\sim (\psi_1 \wedge \psi_2)_{\xi\cdot\chi}}\, \mathrm{I}\wedge$$

$$\frac{\Gamma^W \ |\sim (\psi_1 \wedge \psi_2)_\xi \qquad \Gamma^W \ |\sim \psi_{1\chi}}{\Gamma^W \ |\sim \psi_{2(\xi/\chi)}}\, \mathrm{E}\wedge_R$$
$$\frac{\Gamma^W \ |\sim (\psi_1 \wedge \psi_2)_\xi \qquad \Gamma^W \ |\sim \psi_{2\chi}}{\Gamma^W \ |\sim \psi_{1(\xi/\chi)}}\, \mathrm{E}\wedge_L$$

$$\frac{\Gamma^W \ |\sim \psi_{1\xi} \qquad \Gamma^W \ |\sim \psi_{2\chi}}{\Gamma^W \ |\sim (\psi_1 \vee \psi_2)_{(\xi+\chi)}}\, \mathrm{I}\vee$$

$$\frac{\Gamma^W \ |\sim (\psi_1 \vee \psi_2)_\xi \qquad \Gamma^W \ |\sim \psi_{1\chi}}{\Gamma^W \ |\sim \psi_{2(\xi-\chi)}}\, \mathrm{E}\vee_R$$
$$\frac{\Gamma^W \ |\sim (\psi_1 \vee \psi_2)_\xi \qquad \Gamma^W \ |\sim \psi_{2\chi}}{\Gamma^W \ |\sim \psi_{1(\xi-\chi)}}\, \mathrm{E}\vee_L$$

$$\frac{\Gamma^W, \phi \ |\sim \psi_\xi}{\Gamma^W \ |\sim (\phi \to \psi)_\xi}\, \mathrm{I}\to \qquad \frac{\Delta^W \ |\sim \phi \to \psi_\xi \qquad \Gamma^{W'} \ |\sim \phi_\chi}{\{\Delta,\Gamma\}^{W+W'} \ |\sim \psi_{(\chi\cdot\xi)}}\, \mathrm{E}\to$$

Fig. 5. Rules for Classification.

- $\Gamma^{test,W} := \{P_1(a), \ldots, P_n(a)\}$ is a list of atoms $P_i \in \mathbb{P}^{test}$ for $i = 1, \ldots, n$ over an $a \in \mathbb{D}^{test}$ as per Definition 7 built according to the rules of Fig. 4, and with overall weight W;
- $|\sim^{trained}$ is a non-monotonic inference relation which refers to a (known or unknown) training sample as per Definition 6, and whose semantics is defined by inference rules presented below in Fig. 5;
- ψ is a predicative atomic formula $T(a)$, where $T \in \mathbb{T}$ is the target feature, and $a \in \mathbb{D}^{test}$. Therefore, it is the actual prediction returned by the model for input a;
- $|\xi| \in [0, 1]$ denotes probability that $|\sim^{trained}$ assigns to ψ being validly inferred from $\Gamma^{test,W}$, as determined by the inference rules. Prediction ψ_ξ is returned as a binary classification if we fix a threshold value τ, such that the final output is $T(a)$ if $\frac{\xi+1}{2} \geq \tau$ and is $\neg T(a)$ otherwise. This, in fact, allows for mapping the values in the [-1,1] interval onto probabilities.

Note, importantly, that we will assume function ξ to be identical to function W:

$$\xi = id(W) \tag{7}$$

The rules in Fig. 5 define derivability relations for our classifier. Moreover, a structural rule called (*correction*) is defined to update a prediction's probability as new information (that possibly allows for deriving contradictory conclusions) is added to the original set of observations. Namely, if we combine observations Δ with weights W inferring conclusion ψ with another set of observations Γ with weights W' and inferring conclusion $\neg\psi$, the following holds:

$$\frac{\Delta^W \ |\sim \psi_\xi \qquad \Gamma^{W'} \ |\sim \neg\psi_{\xi'}}{\{\Delta,\Gamma\}^{W+W'} \ |\sim \neg\psi_{(\xi'-\xi)}}\, correction \tag{8}$$

In the calculus, the maximum probability $\xi = 1$ corresponds to the situation in which the inference is based on an amount of information sufficient to infer a conclusion that is *correct*, that is, to infer the ground truth with certainty. In terms of the semantics of the Ulam game, this inference corresponds to the situation in which Responder has used all m available lies. In fact, from this moment onwards, conclusions can be classically entailed on the basis of any new (certainly truthful) answer. In other words, in the game-theoretic semantics, function ξ is the difference between the sum of weighted rejection degrees of those observations that 'speak' against label ψ and the sum of weighted rejection degrees of those observations that 'speak' in favour of it. We capture this intuition more rigorously with the three following definitions.

Definition 12 (gUMO-Norm)**.** Whenever $\Delta \vdash_{RJ} \psi$ satisfies $(\text{gUMO})^j$, we say that $\Delta \vdash_{RJ} \psi$ satisfies $(\text{gUMO-Norm})^{\tilde{j}}$, where:

$$\tilde{j} := \frac{j}{\sum_{H \in \mathcal{H}_\Delta^f} r_\Delta^{uw}(H)}$$

and $\mathcal{H}_\Delta^f$ is the set of hypotheses with a finite degree of rejection.

The level of gUMO-Norm tends to 1 as the difference in rejection degrees between the least and the second least rejected hypothesis increases and the number of finitely rejected hypotheses decreases. It is exactly 1 when $\sum_{H \in \mathcal{H}_\Delta^f} r_\Delta^{uw}(H) = j$, i.e., two finitely rejected hypotheses hypotheses are given, one of which is never rejected. Conversely, the level of gUMO-Norm tends to 0 whenever the difference in rejection degrees between the minimally rejected and the second least rejected hypothesis gets smaller and smaller.

Definition 13. Whenever $\Delta \vdash_{RJ} \psi$ satisfies $(\text{gUMO-Norm})^{\tilde{j}}$, we write $\Delta \vdash_{RJ} \psi_{\tilde{j}}$.

Definition 14. Let a be a generic datapoint in $\mathbb{D}^{test}$. We define γ_Δ as the sum of the signed weights of the predicates satisfied by a, and $\overline{\gamma}_\Delta$ as the sum of the magnitudes of the negative contributions of predicates not satisfied by a. Formally:

$$\gamma_\Delta := \sum_{P_i \in \Delta} \tilde{w}_i \tag{9}$$

$$\overline{\gamma}_\Delta := \sum_{\neg P_i \in \Delta} (|\tilde{w}_i|) \tag{10}$$

Note that, by construction, both γ_Δ and $\overline{\gamma}_\Delta$ are nonnegative as they respectively represent the magnitudes of the predicates speaking 'in favour' and 'against' conclusion ψ.

Proposition 2. *Whenever* $\Delta^W \vdash \psi_\xi, \xi = \gamma_\Delta - \overline{\gamma}_\Delta$.

Proof. By (7):

$$\xi = id(W)$$

where by the *(selection)* rule of Fig. 4:

$$W := \sum_{i \in \{1,\dots,n\}} \tilde{w}_i$$

Let us define the set $\mathcal{I} = \{1, \dots, n\}$ of indexes of predicates in $\mathbb{P}^{test} = \{P_1, \dots, P_n\}$. Therefore, we can rewrite ξ as:

$$\xi = \sum_{i \in \mathcal{I}} \tilde{w}_i$$

We can partition $\mathcal{I}$ into two sets: one containing the indexes of predicates satisfied by a, $\mathcal{S} = \{i | P_i(a) \in \Delta\}$ and the other set containing the indexes of predicates not satisfied by a, $\mathcal{U} = \{i | \neg P_i(a) \in \Delta\}$. Then the total signed weight can be written as

$$\xi = \sum_{i \in \mathcal{S}} \tilde{w}_i + \sum_{i \in \mathcal{U}} \tilde{w}_i$$

For $i \in \mathcal{U}$, we have $\tilde{w}_i < 0$, so $|\tilde{w}_i| = -\tilde{w}_i$. Therefore,

$$\sum_{i \in \mathcal{U}} \tilde{w}_i = \sum_{i \in \mathcal{U}} (-|\tilde{w}_i|) = -\sum_{i \in \mathcal{U}} (|\tilde{w}_i|)$$

Since $\mathcal{S} = \{i | P_i(a) \in \Delta\}$, by construction

$$\sum_{i \in \mathcal{S}} \tilde{w}_i = \sum_{P_i \in \Delta} \tilde{w}_i = \gamma_\Delta$$

Similarly, since $\mathcal{U} = \{i | \neg P_i(a) \in \Delta\}$, by construction

$$\sum_{i \in \mathcal{U}} |\tilde{w}_i| = -\sum_{\neg P_i \in \Delta} (|\tilde{w}_i|) = -\overline{\gamma}_\Delta$$

By Definition 14, they correspond to γ_Δ and $\overline{\gamma}_\Delta$, respectively.
Therefore, substituting this into the expression for

$$\xi = \sum_{i \in \mathcal{S}} \tilde{w}_i + \sum_{i \in \mathcal{U}} \tilde{w}_i$$

we obtain that

$$\xi = \gamma_\Delta + (-\overline{\gamma}_\Delta) = \gamma_\Delta - \overline{\gamma}_\Delta$$

which proves the desired equality. □

Proposition 3. *If* $\mathcal{H} = \{\psi, \neg\psi\}$ *and* $\Delta \vdash_{RJ} \psi$ *satisfies* $(gUMO)^j$*, we say that* $\Delta \vdash_{RJ} \psi$ *satisfies* $(gUMO\text{-}Norm)^{\tilde{j}}$ *where:*

$$\tilde{j} = \frac{r_\Delta^{uw}(\neg\psi) - r_\Delta^{uw}(\psi)}{r_\Delta^{uw}(\neg\psi) + r_\Delta^{uw}(\psi)}$$

In the formula, $r^u w$ *is the Ulam weighted rejection degree as we have defined it earlier in Definition 3.*

Proof.
Since by Definition 13, we know that:

$$\tilde{j} := \frac{j}{\sum_{H \in \mathcal{H}_\Delta^f} r_\Delta^{uw}(H)}$$

we proceed by proving equivalence for the numerator and the denominator of $\tilde{j}$ separately.
(Numerator Equivalence). We want to prove that:

$$j = r_\Delta^{uw}(\neg\psi) - r_\Delta^{uw}(\psi)$$

If $\Delta \nvdash_{RJ} \psi$, then $\hat{\mathcal{H}}_\Delta = \{\psi\}$ by (1) and $\mathcal{H} \setminus \hat{\mathcal{H}}_\Delta = \{\neg\psi\}$. Since $\Delta \nvdash_{RJ} \psi$ satisfies (gUMO)^j, by Definition 4, it follows that $j = \min\{r_\Delta(H') - r_\Delta(H) \mid H \in \hat{\mathcal{H}}_\Delta, H' \in \mathcal{H} \setminus \hat{\mathcal{H}}_\Delta\} = r_\Delta^{uw}(\neg\psi) - r_\Delta^{uw}(\psi)$.

(Denominator Equivalence). We want to prove that:

$$\sum_{H \in \mathcal{H}_\Delta^f} = r_\Delta^{uw}(\neg\psi) + r_\Delta^{uw}(\psi)$$

Assuming that neither hypothesis has been definitely rejected, $\mathcal{H}_\Delta^f = \{\neg\psi, \psi\}$ and therefore $\sum_{H \in \mathcal{H}_\Delta^f} r_\Delta^{uw}(H) = r_\Delta^{uw}(\neg\psi) + r_\Delta^{uw}(\psi)$. □

For convenience, we introduce the following notation:

$$\xi_{\Delta-Bin} := \frac{r_\Delta^{uw}(\neg\psi) - r_\Delta^{uw}(\psi)}{r_\Delta^{uw}(\neg\psi) + r_\Delta^{uw}(\psi)} \tag{11}$$

We introduce the following correspondence.

Proposition 4. *The difference between γ_Δ and $\overline{\gamma}_\Delta$ is equal to $\xi_{\Delta-Bin}$:*

$$\gamma_\Delta - \overline{\gamma}_\Delta = \xi_{\Delta-Bin}$$

Proof. (Proof of Proposition 4)
Given γ_Δ and $\overline{\gamma}_\Delta$, by Proposition 2, we know that the following inference holds in $\nvdash$:

$$\Delta^W \nvdash \psi_{(\gamma_\Delta - \overline{\gamma}_\Delta)}$$

Given the structural rule (*correction*) for the $\nvdash$ derivability relation (8), it is always possible to construct two sets of observations $\Theta^{W''}$ and $\Gamma^{W'}$ such that $\Theta^{W''}$ only contains predicates not satisfied by datapoint a and $\Gamma^{W'}$ only contains predicates satisfied by the same data instance. Note that in this case $\Theta \cap \Gamma = \emptyset$ because if a predicate belongs to the set of predicates satisfied by a, it cannot also belong to the set of predicates not satisfied by a. Let us denote their union by Δ^W, where $\Delta = \Theta \cup \Gamma$ and $W = W'' + W'$.

$$\frac{\Theta^{W''} \nvdash \neg\psi_{\overline{\gamma}_\Delta} \qquad \Gamma^{W'} \nvdash \psi_{\gamma_\Delta}}{\{\Theta, \Gamma\}^{W''+W'} \nvdash \psi_{(\gamma_\Delta - \overline{\gamma}_\Delta)}} \; correction$$

The overall weights W'' and W' are the algebraic sums of the normalised weights associated with predicates in Θ and Γ, respectively. By Definition 11:

$$W' = \gamma_\Delta$$

and similarly:

$$W'' = \overline{\gamma}_\Delta$$

By Definition 9, their absolute values are determined as follows:

$$|W'| = \sum_{P_i \in \Delta} \frac{ww(P_i)}{\sum_{Q \in \Delta} ww(Q)} = \frac{\sum_{P_i \in \Delta} ww(P_i)}{\sum_{Q \in \Delta} ww(Q)}$$

and

$$|W''| = \sum_{\neg P_k \in \Delta} \frac{ww(P_k)}{\sum_{Q \in \Delta} ww(Q)} = \frac{\sum_{\neg P_k \in \Delta} ww(P_k)}{\sum_{Q \in \Delta} ww(Q)}$$

where

$$\sum_{Q \in \Delta} ww(Q) = \sum_{P_i \in \Delta} ww(P_i) + \sum_{\neg P_k \in \Delta} ww(P_k)$$

Since this is exactly how we defined the Ulam weighted rejection degree in Definition 3, we can now introduce the following two equivalences:

$$\sum_{P_i \in \Delta} ww(P_i) = r_\Delta^{uw}(\neg\psi)$$

and

$$\sum_{\neg P_k \in \Delta} ww(P_k) = r^{ww}_{\Delta}(\psi)$$

Therefore:

$$W' - W'' = \frac{r^{ww}_{\Delta}(\neg\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)} - \frac{r^{ww}_{\Delta}(\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)}$$

Finally:

$$\frac{r^{ww}_{\Delta}(\neg\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)} - \frac{r^{ww}_{\Delta}(\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)} = \frac{r^{ww}_{\Delta}(\neg\psi) - r^{ww}_{\Delta}(\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)}$$

which is equal to $\xi_{Bin-\Delta}$ by (11). □

We can now prove the correspondence between the proof-theoretic derivability relation $\vdash$ and the semantic RJ-consequence relation $\vdash_{RJ}$ in the binary case $\{\psi, \neg\psi\}$.

Theorem 1. *$\Delta \vdash \psi_\xi$ with $\xi > 0$ if and only if $\Delta \vdash_{RJ} \psi_\xi$.*

Proof. (Proof of Theorem 1) We show both directions of the equivalence.
(⇒) **From $\vdash$ to $\vdash_{RJ}$:**

Assume $\Delta \vdash \psi_\xi$.
By Proposition 2, we have:

$$\xi = \gamma_\Delta - \overline{\gamma}_\Delta$$

By Proposition 4, we know that in the binary case:

$$\gamma_\Delta - \overline{\gamma}_\Delta = \xi_{\Delta-Bin}$$

By (11), we now that:

$$\xi_{\Delta-Bin} = \frac{r^{ww}_{\Delta}(\neg\psi) - r^{ww}_{\Delta}(\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)}$$

Since $\xi > 0$, then $r^{ww}_{\Delta}(\neg\psi) > r^{ww}_{\Delta}(\psi)$.
Since by assumption $\mathcal{H} = \{\psi, \neg\psi\}$, then $\hat{\mathcal{H}} = \{\psi\}$ and $\mathcal{H} \setminus \hat{\mathcal{H}}_\Delta = \{\neg\psi\}$ so:

$$r^{ww}_{\Delta}(\neg\psi) - r^{ww}_{\Delta}(\psi) = \min\{r_\Delta(H') - r_\Delta(H) \mid H \in \hat{\mathcal{H}}_\Delta, H' \in \mathcal{H} \setminus \hat{\mathcal{H}}_\Delta\}$$

Moreover, by Definition 2, it follows that $\Delta \vdash_{RJ} \psi$.
By Definition 4, $r^{ww}_{\Delta}(\neg\psi) - r^{ww}_{\Delta}(\psi)$ is the (gUMO) level satisfied by $\Delta \vdash_{RJ} \psi$.
We also know that:

$$r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi) = \sum_{H \in \mathcal{H}^f_\Delta} r^{ww}_{\Delta}(H)$$

By Definition 12, $\Delta \vdash_{RJ} \psi$ therefore satisfies (gUMO-Norm)$^\xi$ which, by Definition 13, can be rewritten as:

$$\Delta \vdash_{RJ} \psi_\xi$$

(⇐) **From $\vdash_{RJ}$ to $\vdash$:**
Assume $\Delta \vdash_{RJ} \psi_\xi$.
By Definition 13, this means that $\Delta^W \vdash_{RJ} \psi$ satisfies (gUMO-Norm)$^\xi$.
By Definition 12, we know that:

$$\xi = \frac{j}{\sum_{H \in \mathcal{H}^f_\Delta} r^{ww}_{\Delta}(H)}$$

By Proposition 3, we know that in the binary case:

$$\frac{j}{\sum_{H \in \mathcal{H}^f_\Delta} r^{ww}_{\Delta}(H)} = \frac{r^{ww}_{\Delta}(\neg\psi) - r^{ww}_{\Delta}(\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)}$$

By (11):

$$\frac{r^{ww}_{\Delta}(\neg\psi) - r^{ww}_{\Delta}(\psi)}{r^{ww}_{\Delta}(\neg\psi) + r^{ww}_{\Delta}(\psi)} = \xi_{\Delta-Bin}$$

By Proposition 4:

$$\xi_{\Delta-Bin} = \gamma_\Delta - \overline{\gamma}_\Delta$$

By Proposition 2, we can formulate the inference:

$$\Delta \mid\sim \psi_{(\gamma_\Delta - \overline{\gamma}_\Delta)}$$

with $\gamma_\Delta - \overline{\gamma}_\Delta > 0$ because $\gamma_\Delta - \overline{\gamma}_\Delta = \xi_{\Delta - Bin} = \xi$ and ξ is positive by assumption.
In light of the same reason, we rewrite it as

$$\Delta \mid\sim \psi_{\xi}$$

□

8. Conclusion and future work

In the present investigation, a variation of the rational non-monotonic consequence relation inspired by the two-player Rényi-Ulam game was used to reason about the impact of data bias on ML outcomes. In particular, we focused on defining a novel way to test whether such outcomes are fair, assuming no causal knowledge on the data-generating process, but only knowing the model's input data and its associated decisions.

Our proposal can be expanded further by exploring more realistic and sophisticated models of measurement errors. One way to do so is to explore another variant of the Ulam game with channels, namely the Ulam game with channels and probabilistic half-lies. In this game setting, whether Responder can lie depends not only on the outcome of an aleatory experiment *but also on the true answer itself*. Half-lies help us to model those scenarios in which the probability of false negatives is significantly higher than that of false positives, or *vice versa*. In the case above of the biased oximeter reading, for example, we would have that Responder tosses a coin biased towards Head with $p = 0.8$ whenever the patient's *true* oxygen level is below the threshold. If Head is returned, then Responder lies and gives back a high oxygen level. In contrast, whenever the oxygen level is above the threshold, no coin is tossed and Responder is forced to return the truthful answer. In other words, while Questioner can trust observations indicating low blood oxygen levels, the same confidence cannot be applied to observations of normal blood oxygen levels. In this version of the game, we never reach complete certainty about the secret but, being the outcome of the aleatory experiments independent, Nature's answers will nevertheless asymptotically converge towards the true hypothesis.

Although the above discussion has mainly focused on the case of measurement error in the medical domain, the general framework is sufficiently flexible to apply to a broader range of cases within fair ML research. Each of these cases may require domain-specific error models and considerations. For instance, the disadvantaged sensitive group is not necessarily the one associated with noisier data; individuals may in fact even be advantaged by noisier channels. Consider, for example, a fraud-risk assessment context in which individuals are evaluated on the basis of features such as past frauds or criminal records. These data may be more accurate for certain sensitive groups – those more frequently scrutinised for fraud or more heavily policed – than for others. Again, probabilistic half-lies can be employed to capture asymmetric informativeness: the presence of a criminal record is a reasonably faithful indicator of past crime, whereas its absence does not necessarily indicate that no crime has occurred. Crucially, the degree of this asymmetry may differ systematically between sensitive groups. In this scenario, individuals associated with the noisier channel end up being *advantaged*, as spotting fraudsters among them becomes a harder task.

Future research will provide an experimental analysis for ML scenarios of this kind.

CRediT authorship contribution statement

Chiara Manganini: Conceptualization, Writing – original draft, Methodology, Formal analysis; **Esther Anna Corsi:** Writing – original draft, Methodology, Formal analysis; **Giuseppe Primiero:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

All authors were supported by the Ministero dell'Università e della Ricerca (MUR) through the Project "Departments of Excellence 2023-2027" awarded to the Department of Philosophy "Piero Martinetti" of the University of Milan. Giuseppe Primiero was additionally supported by MUR through the PRIN 2022 Project "SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin" (grant code 20223E8Y4X). Esther Anna Corsi was additionally supported by the 2023 Young Researchers Postdoctoral Fellowship funded by Fondazione Cariplo, under the Project "Logics for Scientific Inferences (LOGSI)" (grant code 2023-0978).

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, 2022, 1908.09635.
[2] S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im)possibility of fairness, 2016, 1609.07236.
[3] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, K.R. Varshney, Is there a trade-Off between fairness and accuracy? a perspective using mismatched hypothesis testing, in: International Conference on Machine Learning, 2020. https://api.semanticscholar.org/CorpusID:220544367.

[4] A.Z. Jacobs, H. Wallach, Measurement and fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 375–385. https://doi.org/10.1145/3442188.3445901.
[5] S.M. Ulam, Adventures of a mathematician, in: Mathematics, Chapman and Hall/CRC, 2019, pp. 193–200.
[6] A. Pelc, Searching games with errors—Fifty years of coping with liars, Theor. Comput. Sci. 270 (1-2) (2002) 71–109. https://doi.org/10.1016/S0304-3975(01)00303-6.
[7] P. Baldi, E.A. Corsi, H. Hosni, A logical framework for data-driven reasoning, Logic J. IGPL (2024). https://doi.org/10.1093/jigpal/jzae113.
[8] T. Rienstra, Explaining boolean classifiers with non-monotonic background theories, in: F. Oliehoek, M. Kok, S. Verwer (Eds.), Artificial Intelligence and Machine Learning - 35Th Benelux Conference, BNAIC/Benelearn 2023, Revised Selected Papers, 2187 CCIS of *Communications in Computer and Information Science*, Springer, United States, 2025, pp. 174–188. Publisher Copyright: © The Author(s), under exclusive license to Springer Nature Switzerland AG 2025.; 35th Benelux Conference on Artificial Intelligence and Machine Learning, BNAIC/Benelearn 2023 ; Conference date: 08-11-2023 Through 10-11-2023, https://doi.org/10.1007/978-3-031-74650-5_10.
[9] M. Whitehead, R. Ali, E. Carrol, C. Holmes, F. Kee, Equity in Medical Devices: Independent Review, Technical Report, 2023. https://assets.publishing.service.gov.uk/media/65e89e9e62ff48001a87b2d8/equity-in-medical-devices-independent-review-report-web-accessible.pdf.
[10] M.W. Sjoding, R.P. Dickson, T.J. Iwashyna, S.E. Gay, T.S. Valley, Racial bias in pulse oximetry measurement, N. Engl. J. Med. 383 (25) (2020) 2477–2478.
[11] A. Fawzy, T.D. Wu, K. Wang, M.L. Robinson, J. Farha, A. Bradke, S.H. Golden, Y. Xu, B.T. Garibaldi, Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19, JAMA Intern. Med. 182 (7) (2022) 730–738.
[12] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
[13] F. Cicalese, D. Mundici, Recent developments of feedback coding and its relations with many-valued logic, Proof, Comput. Agency: Logic Crossroads (2011) 115–131.
[14] F. Montagna, C. Marini, G. Simi, Product logic and probabilistic ulam games, Fuzzy Sets Syst. 158 (6) (2007) 639–651. The Logic of Soft Computing, https://doi.org/10.1016/j.fss.2006.11.007.
[15] E.A. Corsi, F. Montagna, The rényi–Ulam games and many-valued logics, Fuzzy Sets Syst. 301 (2016) 37–50.
[16] A. Noriega-Campero, M.A. Bakker, B. Garcia-Bulle, A.S. Pentland, Active fairness in algorithmic decision making, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2018). https://api.semanticscholar.org/CorpusID:52896535.
[17] M.A. Bakker, H.R. Valdés, D.P. Tu, K.P. Gummadi, K.R. Varshney, A. Weller, A.S. Pentland, Fair enough: improving fairness in budget-Constrained decision making using confidence thresholds, in: SafeAI@AAAI, 2020. https://api.semanticscholar.org/CorpusID:211840427.
[18] A. Rényi, On a problem of information theory, A MAGYAR TUDOMÁNYOS AKADÉMIA MATEMATIKAI KUTATÓ INTÉZETÉNEK KÖZLEMÉNYEI 6 (4) (1961) 515–516.
[19] D. Mundici, et al., The logic of Ulam's game with lies, in: Knowledge, Belief and Strategic Interaction, Cambridge Studies in Probability, Induction, and Decision Theory, Cambridge University Press, 1992, pp. 275–284.
[20] D. Mundici, Ulam games, Łukasiewicz logic, and AF c*-Algebras, Fundam. Inform. 18 (2–4) (1993) 151–161.
[21] F. Cicalese, Fault-Tolerant Search Algorithms: Reliable Computation with Unreliable Information, Springer Publishing Company, Incorporated, 2013.
[22] Q. Kang, W.P. Tay, Sequential multi-Class labeling in crowdsourcing: a ulam-Renyi game approach, in: Proceedings of the International Conference on Web Intelligence, WI '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 245–251. https://doi.org/10.1145/3106426.3106446.
[23] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 214–226. https://doi.org/10.1145/2090236.2090255.
[24] S. Barocas, A.D. Selbst, Big Data's disparate impact, Calif. Law Rev. 104 (2016) 671.
[25] M.J. Kusner, J.R. Loftus, C. Russell, R. Silva, Counterfactual Fairness, 2018, 1703.06856.
[26] S. Venkatasubramanian, M. Alfano, The philosophical basis of algorithmic recourse, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 284–293. https://doi.org/10.1145/3351095.3372876.
[27] A.-H. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 353–362. https://doi.org/10.1145/3442188.3445899.
[28] B. Ustun, A. Spangher, Y. Liu, Actionable recourse in linear classification, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 10–19. https://doi.org/10.1145/3287560.3287566.
[29] J. von Kügelgen, U. Bhatt, A.-H. Karimi, I. Valera, A. Weller, B. Scholkopf, On the Fairness of Causal Algorithmic Recourse, ArXiv abs/2010.06529 (2020). https://api.semanticscholar.org/CorpusID:222310477.
[30] E. Sullivan, A. Kasirzadeh, Explanation Hacking: The Perils of Algorithmic Recourse, in: J.M. Durán, G. Pozzi (Eds.), Philosophy of science for machine learning: Core issues and new perspectives, Springer, forthcoming.
[31] M. Cerrato, A. Vallenas Coronel, M. Köppel, The case for correctability in fair machine learning, in: Proceedings of the European Workshop on Algorithmic Fairness (EWAF'23), Winterthur, Switzerland, 2023, pp. 1–4.
[32] C. Manganini, G. Primiero, Reasoning With and About Bias, Springer Nature Switzerland, Cham, 2024, pp. 127–154. https://doi.org/10.1007/978-3-031-77892-6_7.