# A Comprehensive Ethical Analysis on Societal Proxies and Algorithmic Fairness

Matthias Bartolo
*matthias.bartolo.21@um.edu.mt*
*University of Malta*
Msida, Malta

Jerome Agius
*jerome.agius.21@um.edu.mt*
*University of Malta*
Msida, Malta

Isaac Muscat
*isaac.muscat.21@um.edu.mt*
*University of Malta*
Msida, Malta

David Cachia Enriquez
*david.cachia-enriquez.21@um.edu.mt*
*University of Malta*
Msida, Malta

Matthew Kenely
*matthew.kenely.21@um.edu.mt*
*University of Malta*
Msida, Malta

***Abstract*—This paper provides a comprehensive introduction to the concepts of proxies and algorithmic fairness, as well as an ethical analysis of both. Additionally, this research outlines a variety of proxies, making reference to formal mathematical representations as well as societal applications, in addition to providing potential reasons for their application. In the case of misused proxies, the potential for discrimination, both causal and opaque, is underlined through the analysis of case studies and formalisation with respect to the machine learning process. An in-depth definition of algorithmic fairness and the challenges it poses are provided. This research highlights different forms of bias in AI models and their causes, in conjunction with methods for mitigating them. An ethical discussion on algorithmic decision-making is presented, with an emphasis on the degree of impact that these algorithms have on individuals as reflected in the EU AI Act. The paramount responsibility of dataset curators and machine learning engineers in creating unbiased datasets and continuously monitoring AI models is highlighted. The advantages and shortcomings of both AI models and subjective human judgement are compared and contrasted, and an approach incorporating a balance of both moving forward is proposed.**

***Index Terms*—proxy, proxy discrimination, algorithmic fairness, algorithmic decision-making, ethical artificial-intelligence, dataset bias**

## I. Introduction

### A. Proxy Definition

The word “proxy” stems from the noun “procuracy”, i.e., the position of one who carries out an action in another’s stead or on their behalf. The concept of a proxy is similar to that of a vicar in the Roman Catholic Church, who represents a bishop, a role that led to the adoption of the idea of something being carried out vicariously, i.e., for someone else [1].

In general, a proxy is an entity that can both stand for (represent) and stand in for (replace) another item. Mathematically, a proxy can be represented using Implication 1, which states that $P$ is the proxy for $R$ if and only if $P$ can both represent and replace $R$.

$$\text{Proxy}(P,R) \implies \text{Represents}(P,R) \land \text{Replaces}(P,R) \tag{1}$$

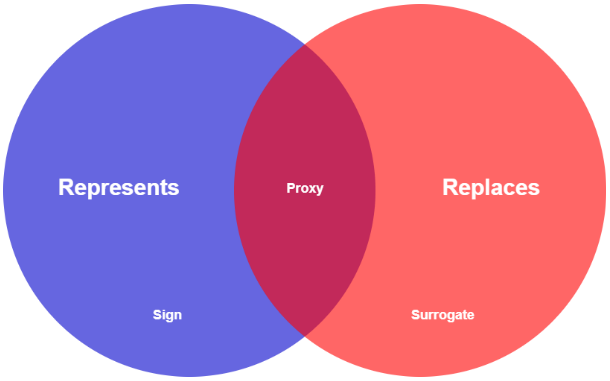


Fig. 1. Venn Diagram showcasing the different types of proxies according to [1].

### B. Types of Proxies

Floridi [1] introduces the concept of *degenerate proxies*, which imply the limitation of an object to belonging to a particular class, usually simpler. He provides the example of a point being a degenerate sphere (a sphere of radius zero). Degenerate proxies can be divided into two categories: these being *Signs* and *Surrogates*.

*Signs* refer to a type of degenerate proxy in which the proxy $P$ represents $R$ but does not replace it, as stated in Implication 2.

$$\text{Sign}(P,R) \implies \text{Represent}(P,R) \land \neg\text{Replaces}(P,R) \tag{2}$$

Moreover, signs can be further categorised into three subcategories: icons, indexes, and symbols.

- Icons are a type of sign that resemble what they stand for; for instance, an image of a cloud represents said cloud.
- Indexes are signs that correlate with what they represent; for instance, a cloud represents the impeding rain.
- Symbols are a type of sign that represent an object by virtue of some convention; for example, the word cloud

is used to represent a mass of floating vapour in an atmosphere.

*Surrogates* refer to a type of degenerate proxy in which the proxy $P$ replaces $R$ but does not represent it, as stated in Implication 3.

$$\text{Surrogate}(P, R) \implies \neg\text{Represent}(P, R) \wedge \text{Replaces}(P, R) \quad (3)$$

### C. Application of Proxies

Intrinsically, proxies are used to represent some variable of interest for three main reasons:

1) When the variable of interest could not be otherwise measured directly.
2) When the variable of interest is too costly to measure in terms of time or money.
3) When the variable of interest has a lot of missing values and thus cannot be used in statistical analysis.

Instances of these variables include intelligence and quality of life, which cannot be meaningfully represented numerically. Proxies have been used to bridge this statistical gap and represent both of these variables (IQ and GDP per capita, respectively) [2].

For a proxy to be a good representation and replacement of some variable of interest, it must satisfy two requirements: high predictability in relation to the represented variable and a lack of bias [3]. The relationship between proxies and the variables they represent does not necessarily need to be linear, but an accurate prediction of variables should be achievable by utilising proxies.

With respect to mitigating bias, a good proxy can contain variability and noise and does not need to employ the same values used by the original variable. Thus, a good proxy is denoted by the lack of bias towards a particular trait or property of the original variable [4].

## II. Understanding Societal Proxies

*Societal proxies* are abstract representations that portray socially salient information regarding the presence and activities of a group of people. They can exist both online and offline; the latter includes traditional media[1], public demonstrations, surveys, and interviews. On the contrary, online proxies include social media trends, online communities, and news websites. These proxies serve a crucial role in society as they provide a means by which people can understand society better. For instance, the monitoring of these societal proxies allows researchers, businesses, and policymakers to study public preferences, determine market trends, and identify public opinion, respectively [5].

Societal proxies can take various forms amongst a variety of different groups throughout society, as can be seen in Figure 2. These include:

### A. Educational Proxies

Educational proxies can take various forms. For example, they may be integrated into the teaching process to represent larger concepts, phenomena, or variables. Additionally, external tutors and mentors can be seen as proxies for the institution or educators as they guide the students in their studies. A student's emotional, behavioural, and cognitive engagement can serve as a proxy for how well they will perform when assessed [6]. Alternatively, the authors in [7] mentioned the use of *Grade Point Average (GPA)*, *Scholastic Aptitude Test (SAT)*, degrees, and diplomas as proxies for a student's success in their scholastic endeavours. Moreover, educational institutions may utilise external proxies such as skin colour or financial stability as an indication for a student requiring extra assistance. The first three proxies pose no apparent issues, whereas the latter two can affect which students are provided with extra opportunities and which are afforded extra help. This may result in bias, which hinders students who require help but, according to the proxy, are not deemed to need it [8].

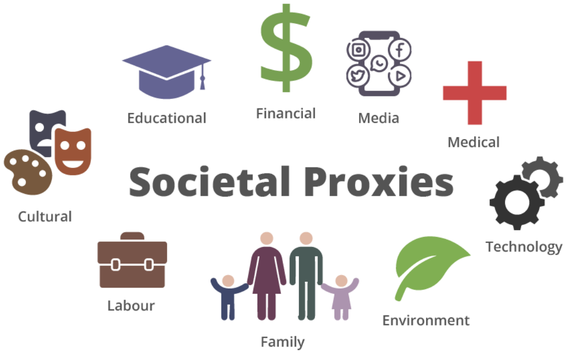


Fig. 2. Types of Societal Proxies.

### B. Financial Proxies

Financial proxies embody various implementations. They can be incorporated into methodologies, such as the *Social Return on Investment (SROI)*, to quantify the monetary value of social, environmental, and other impacts that extend beyond traditional success metrics. In the context of SROI, these financial proxies serve as crucial components, acting as substitutes for assigning value to specific impacts. For instance, Canada's Financial Proxies Database provides practitioners with a comprehensive set of indicators for calculating SROI ratios. These proxies, derived from reliable financial sources, are indispensable for articulating the maximum value generated and enhancing the credibility of the analysis and report [9].

Further expanding on financial proxies, *Gross Domestic Product (GDP)* and *credit scores* serve as additional tools to assess economic impact. GDP serves as a proxy reflecting the overall economic well-being of a country or region [10], [11],

[1] Television, radio, newspapers

while credit score represents an individual's financial stability and responsibility [12].

Financial proxies play a pivotal role in conveying the significance of an organisation's efforts, supporting evidence-based decision-making, and optimising resource allocation. It is essential to ensure the accuracy of financial proxies to capture the true economic and social contributions of organisations, fostering a more informed understanding of their impact on communities [9].

### C. Cultural Proxies

Cultural proxies encompass the different traits, characteristics, and aspects related to a particular culture that act as stand-ins for broader cultural diversity. Such types of proxies can be applied to multiple different sectors to increase diversity, including educational, professional, and even social settings.

An example of a cultural proxy in education would be to consider the race and religion of an individual as a proxy to achieve diversity goals. In [13], the author emphasises the significance of using race or religion as a cultural proxy in education. Many schools tend to lack racial and cultural diversity because of various factors that have perpetuated inequality in access to education. To address this issue, incorporating cultural proxies such as race and religion can serve as a crucial tool to ensure a more diverse learning environment.

Including religion as a cultural proxy for diversity could, however, introduce fraudulent cases where people misrepresent a religion to gain a better advantage for diversity selection. On the other hand, the exclusion of religion as a cultural proxy from diversity considerations raises questions regarding the narrow tailoring of such programs. The inherent challenges of sorting people based on cultural attributes, such as race or religion, become apparent when the intention is to achieve cultural and intellectual diversity for legitimate purposes.

In [14], the authors delve into the persistent debate within genetics regarding the biological significance of categorising race and ethnicity. They also address the associated challenges and emphasise the need for researchers to justify the scientific utility of employing social identities in genomic[2] research. This shows that, despite the intention of cultural proxies, their use may pose difficulties in aligning the means with the desired outcomes, requiring careful consideration and ongoing refinement.

### D. Medical Proxies

Medical proxies encapsulate a variety of aspects used throughout the medical field. This includes *Durable Power of Attorney for Health Care (DPAHC)* and *Body Mass Index (BMI)*. The DPAHC is a legal arrangement that allows an individual (proxy) to make decisions on behalf of another, given that they are unable to do so themselves. Its main aim is to aid older adults in sparing their families from difficult disagreements over life and death decisions by appointing an individual to make said choices. This proxy is available to the public through an opt-in system [15]. Several aspects are taken into consideration when a person appoints a DPAHC these include family roles, relationships, socioeconomic status, health, personal beliefs, and direct experiences with end-of-life issues [15]. BMI can be used as a proxy for an individual's health, as it incorporates weight with respect to height, which may be roughly indicative of a person's body fat percentage. However, due to how BMI is calculated, it may not necessarily be an accurate representation of a person's health in all cases [16].

### E. Family Proxies

Family proxies refer to individuals who serve as substitutes for someone else, usually with a familial connection. They can make decisions or communicate information on their behalf in the event that the suspect individual is not able to communicate or convey their thoughts accurately. These substitutes are responsible for describing the feelings, preferences, or thoughts of the suspect individual, who may be facing challenges in self-expression for various reasons, such as health conditions, cognitive impairments, or language barriers.

A study conducted by Heid *et al.* [17] found that family members accurately reported the preferences of nursing home residents. However, a few disagreements were found to be mainly related to personal development-related activities. Factors such as the individual's perceptions of the resident's traits influenced these discrepancies. Similarly, in another study carried out by Williams *et al.* [18], it was observed that stroke survivors and their family proxies only moderately agreed on the overall *Health-Related Quality of Life (HRQL)* two months after the stroke. The family proxies tended to rate the HRQL lower than the survivors, particularly in subjective topics such as mood.

These studies emphasise the importance that family proxies play in healthcare research and decision-making processes. However, it may also be the case that family proxies do not accurately represent the individual's preferences, values, or HRQL. This can occur due to underestimations of the importance or discrepancies in the understanding of the personal characteristics of the individual. Hence, while family proxies are valuable in scenarios where the suspect patient cannot provide information themselves, it is essential to approach these proxies with the recognition that they may have limitations.

### F. Environment Proxies

Environmental proxies are indicators or metrics used in scientific research to infer and understand our environment. These proxies, which can serve as substitutes for biological, chemical, or geological markers, provide insights into the past and present conditions of the environment. In a study [19], otolith chemistry is used as an environmental proxy in fisheries research. These offer valuable insights into long-term environmental changes and the effects of human interaction with aquatic environments.

The authors in [20] explore the use of environmental footprints as a proxy for comparing the environmental impact

[2] The study of genetic and epigenetic information in organisms.

caused by some entities, ranging from a product to an entire nation. This study underscores the importance of environmental footprints as a proxy for human health and biodiversity. Another study [21] used isotope ratios from plant waxes within lake cores in Cameroon to understand deforestation. Using proxies such as this helps with comprehending and uncovering the history of an environment and the effects of human interaction over the years. However, using environmental proxies entails inherent risks, including the risk of misinterpretation due to the complexity of natural systems and human interactions. Factors such as spatial and temporal variability, as well as missing gaps in data, contribute to uncertainties in the reconstruction of the environment. Additionally, the reliance on proxies introduces the risk of bias, as the selected proxies may not fully capture the dynamics of past environmental conditions, emphasising the need for cautious interpretation of proxy-based studies.

### G. Labour Proxies

Labour or employment proxies are crucial representatives, acting as intermediaries between companies and prospective employees, going beyond just matching skills. However, relying upon proxy-reported US data, particularly for gender, ethnicity, and disability proxies, can significantly impact the understanding of economic frameworks. As reported in [22], proxy bias in disability reporting arises from inconsistencies between self-reports and proxy reports, notably among individuals aged between 18 and 64. Proxies (individuals) tend to under-report impairments, thereby influencing the accuracy of national figures.

According to [23], an examination of the *Current Population Survey (CPS)* reveals that self-reported salaries outnumber proxy-reported wages, influencing estimates of the gender wage gap. This disparity highlights the complexities of reporting biases affected by cultural standards. Furthermore, the use of proxies in surveys like the CPS is motivated by financial concerns, yet it has a substantial impact on society's views of labour market trends. On the other hand, as pointed out by [24], the *Consumer Price Index (CPI)*[3] informs economic policy choices. Combining self and proxy responses necessitates careful thought, particularly when examining historical patterns or gender-based analysis. Critical examination becomes crucial when investigating proxies such as the "reasonable person" in legal frameworks or economic indices such as CPI, due to their tremendous influence on social views and regulatory actions.

### H. Media Proxies

Media proxies typically act as intermediaries between users and media sources, and often facilitate access or content delivery. In the study [25], several media tools, such as *Google Trends*, *Instagram* and *Twitter* (now *X*) served as proxies for evaluating social distancing measures during the COVID-19 pandemic in the US. Additionally, the study identified an inverse correlation between social media activity and the *time-varying reproduction number (Rt)*. Rt serves as an epidemiological estimate of the virus' transmission rate. Social media served to showcase a weaker but earlier correlation with Rt compared to social mobility measures, indicating how social media can serve as an early indicator of future social behaviour. Similarly, [26] makes use of geo-tagged tweets as a proxy for modelling hourly electric power consumption at the building level.

According to the findings of these studies, Twitter human activity indicators serve as a good correlation (0.8) for power consumption. Proxies such as these could lead to better power consumption demand predictions being carried out. Finally, another study [27] discusses the use of the *bounce rate* as a proxy relating to *click-through rate (CTR)* and *conversion rate (CvR)*. In the research, bounce rate denotes the percentage of visitors to a particular website who navigate away from the site after viewing only one page. The findings showcase an inverse relationship between bounce rate and CTR/CvR. Although these studies outline the positive applications of media proxies, they are accompanied by a variety of disadvantages. These include the perpetuation of bias due to media proxies, which might incorrectly represent groups of people, and privacy issues, which are a primary concern where personal user data may be considered.

### I. Technology Proxies

Technological proxies act as indicators in scientific research to represent the influence of technology on various phenomena. These proxies allow researchers to examine and measure the effect of technological factors such as internet usage or neural network models on outcomes including wages, enhanced oil recovery, or social interactions. In one study where internet usage was employed as a proxy for IT skills in a workplace [28], logit models were utilised to estimate the impact of this technology on wages. The findings demonstrated positive and statistically significant effects, indicating higher wages for internet users. The extent of these wage differences ranged from 4.9% to 16.4%, depending on the level of technological requirements in different industries.

In a different study on enhanced oil recovery [29], artificial neural network models were implemented as proxies to predict oil production rates. These models efficiently estimate the impact of various engineering parameters on water huff-n-puff technology for the recovery of oil in damaged reservoirs.

In a study examining technology as a proxy [30], the research emphasised the role of technology like the *Web2gether* system in influencing social interactions and cognitive processes. This approach highlights the challenges in human-computer interaction, necessitating a balance between traditional usability studies and real-world effectiveness, considering social practices, norms, and cultures alongside technological changes.

Despite the multitude of usages technology proxies can find themselves in, they also introduce inherent risks, including their possible misinterpretation due to the complexity

[3] A key indicator that tracks changes in consumer expenditures

of technological systems or their dynamic interactions. The ever-changing nature of technology, accompanied by ethical considerations and bias, amplify the uncertainties of using technological proxies. This further emphasises the importance of interpreting results with caution when produced via proxy-based work.

While societal proxies play a pivotal role in creating correlations to better understand and interact with the environment around us, it is vital that the people who make use of them recognise their limitations and possible risks. As discussed prior, these proxies can be implemented in multiple sectors. However, they also risk perpetuating social bias. *Social Bias* is the unfair discrimination based on stereotyping, favouring one group of individuals over another, prevalent in racism and misogyny. The misinterpretation of proxies can create biases and inaccuracies, compromising the reliability of conclusions drawn from studies that make use of them. To cater for this, critical examination and awareness of these potential biases are essential, ensuring accurate interpretation of these conclusions [31].

## III. Proxy Discrimination

Proxy discrimination is a phenomenon seen in artificial intelligence and big data where the use of a proxy proves harmful to members of a certain class. Proxy discrimination can be both *intentional* and *non-intentional*.

*Intentional* proxy discrimination occurs when one knowingly uses a proxy to replace some variable of interest that is legally prohibited from being discriminated against.

*Unintentional* proxy discrimination occurs when membership in a class is correlated with the discriminator's neutral goal[4], thus making this type of discrimination rational [3]. An example of unintentional proxy discrimination is when an algorithm uses a person's music interests as a proxy for their race, potentially perpetuating racial stereotypes. Proxy creation may combine multiple characteristics, possibly discriminating against a particular class of members [4].

A real-life example of proxy discrimination occurred at the beginning of the 21st century, when the *U.S. Equal Employment Opportunity Commission (EEOC)* investigated a case related to the use of pre-employment testing when selecting which employees to hire. The company in question, *Target*, designed an algorithm meant to uniformly determine the employees' cognitive abilities and skills. Research showed, however, that African Americans and Latinos in particular were unfairly discriminated against, thereby negatively impacting their job prospects. This proxy discrimination, although unintentional, occurred due to the algorithm's disparate impact on the aforementioned minority groups [32].

### A. Proxy Discrimination in Artificial Intelligence

Artificial intelligence (AI) requires training data to build a model. The AI extracts features from the training data to create complex models that connect the input data to the target variable. Unlike other models, AI does not use any pre-existing hypotheses. Instead, it makes use of all the available characteristics to maximise the desired outcome. As a result, the statistical models generated by AI can be difficult for programmers to explain, and since they work independently from the programmers, they may seem resistant to engaging in intentional discrimination. Regardless, there is a risk of proxy discrimination occurring within AI models. Due to legal bindings, certain characteristics are omitted from the training data, with the aim of reducing discrimination. However, this results in less obvious proxies being discovered and utilised as substitutes by the AI models.

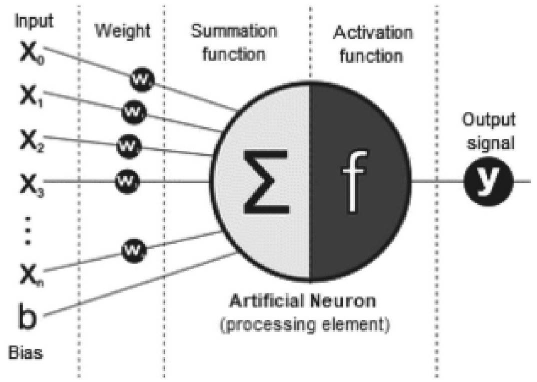


Fig. 3. Schematic representation of the mathematical model of an artificial neuron (processing element), highlighting input ($x_i$), weight ($w$), bias ($b$), summation function ($\Sigma$), activation function ($f$) and output signal ($y$). (Source: [33])

As AI and big data continue to advance, proxy discrimination will become increasingly detrimental and a greater challenge for anti-discrimination regimes that attempt to prohibit it. Despite these increasing risks, current laws and policies fail to properly differentiate between proxy discrimination and disparate impact analysis. The lack of research and development carried out in this area could hinder the efficiency of measures taken to reduce this discrimination.

The creation of proxies occurs due to a reduction in feature space as a result of imposed constraints. This reduction requires the model to utilise proxies, as seen in Equation 6, to achieve the target variable, as these omitted features, most commonly consisting of sex, age, race, and disability [3], tend to be conducive to the target variable. The general dataset structure is outlined in Equation 4 where $x$ denotes the features and $y$ the target label.

Following this, the AI model, which comprises a sequence of weights and biases, is trained on the dataset, as illustrated in Equation 5 [33]. Given the reduced feature space, it will substitute the aforementioned features to facilitate prediction of the target variable [3].

Dataset $D$ with $n$ features denoted as $x_1, x_2, \ldots, x_n$ and the target variable $y$:

$$D = \{(x_1, x_2, \ldots, x_n), y\} \tag{4}$$

[4] An objective or intention that, on its surface, is not discriminatory or biased

AI model with weights $w_0, w_1, \ldots$ and biases $b$:
The variable $x$ represents the input features, while $\hat{z}$ denotes the predicted output generated by the AI model as a function of $x$, with weights $w$ and biases $b$:

$$\hat{z} = f(x \mid w, b) \tag{5}$$

If a variable $x_k$ is not available, the model may attempt to substitute it with a (potentially) relevant proxy variable $p_k$

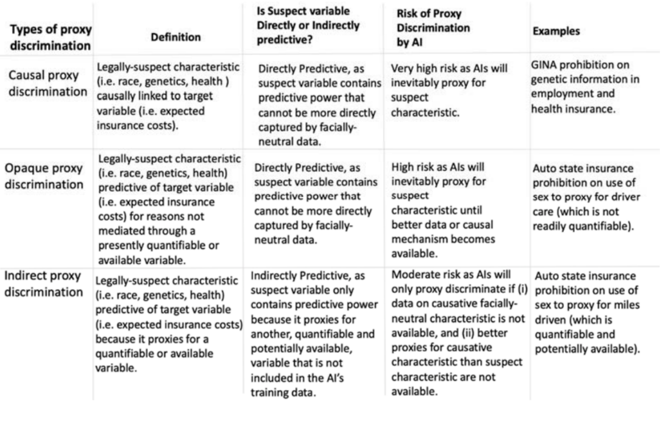
$$\hat{z} = f(p \mid w, b) \tag{6}$$

Anti-discrimination laws are enacted with the aim of prohibiting discrimination through directly predictive traits. However, this negatively affects the anti-discriminatory regimes that are most at risk of AI discrimination (health insurance, non-health insurance, employment, and other legal areas). It undermines their intended goals by promoting social risk sharing and preventing the chilling[5] of socially valuable behaviour, limiting or reversing the effects of past discrimination, and anti-stereotyping.

Proxy discrimination by AI can, however, be alleviated through the following courses of action:

- Prohibiting discrimination based on non-approved factors.
- Mandating the collection and disclosure of data about impacted individuals' membership in legally protected classes.
- Requiring firms to employ ethical algorithms that explicitly control for proxy discrimination.
- Requiring firms to demonstrate potential causal connections between their decision criteria and their legitimate objectives.

Proxy discrimination can take on various forms, these being *causal*, *opaque*, and *indirect* as seen in Figure 4.



| Types of proxy discrimination | Definition | Is Suspect variable Directly or Indirectly predictive? | Risk of Proxy Discrimination by AI | Examples |
|---|---|---|---|---|
| Causal proxy discrimination | Legally-suspect characteristic (i.e. race, genetics, health ) causally linked to target variable (i.e. expected insurance costs). | Directly Predictive, as suspect variable contains predictive power that cannot be more directly captured by facially-neutral data. | Very high risk as AIs will inevitably proxy for suspect characteristic. | GINA prohibition on genetic information in employment and health insurance. |
| Opaque proxy discrimination | Legally-suspect characteristic (i.e. race, genetics, health) predictive of target variable (i.e. expected insurance costs) for reasons not mediated through a presently quantifiable or available variable. | Directly Predictive, as suspect variable contains predictive power that cannot be more directly captured by facially-neutral data. | High risk as AIs will inevitably proxy for suspect characteristic until better data or causal mechanism becomes available. | Auto state insurance prohibition on use of sex to proxy for driver care (which is not readily quantifiable). |
| Indirect proxy discrimination | Legally-suspect characteristic (i.e. race, genetics, health) predictive of target variable (i.e. expected insurance costs) because it proxies for a quantifiable or available variable. | Indirectly Predictive, as suspect variable only contains predictive power because it proxies for another, quantifiable and potentially available, variable that is not included in the AI's training data. | Moderate risk as AIs will only proxy discriminate if (i) data on causative facially-neutral characteristic is not available, and (ii) better proxies for causative characteristic than suspect characteristic are not available. | Auto state insurance prohibition on use of sex to proxy for miles driven (which is quantifiable and potentially available). |

Fig. 4. Types of Proxy Discrimination (Source: [3])

*Causal proxy discrimination* refers to instances in which AI models inadvertently use proxies correlated with protected characteristics (race or gender) to facilitate decision-making or predictions. An example of such a causal relationship is the gene for Huntington's disease. Given that an individual has a series of nucleotide repeats in the Huntington gene over a set threshold, they will always develop the disease, whereas those below the threshold will not. Causal relationships can be hard to isolate; thus, when sophisticated AIs are deprived of this direct information due to legal restrictions, they use any available data (including partial proxies) as a substitute. Given the example of Huntington's disease, assuming the algorithm is prohibited from considering genetic tests for the disease, these tests could proxy for the disease through variables such as family medical history or visits to a Huntington's disease support group website [3].

*Opaque proxy discrimination*, similar to Causal proxy discrimination, makes use of protected characteristics however, these characteristics are not mediated through a presently quantifiable or available variable. Opaque proxy discrimination can occur in two distinct ways. First, this occurs when the initial variable for which a proxy exists cannot be quantified as it is not fully understood. This might occur because the proxy is causative or is standing in for a causative variable. For instance, consider the hiring process. An AI system might use an applicant's university as a proxy variable to predict job performance. However, the specific qualities or characteristics that make a particular university a strong predictor of success in a given role might not be fully known or measurable. The proxy, in this case, could be representing an assortment of factors, such as the quality of education, networking opportunities, or other nuanced elements that contribute to an individual's suitability for the job [3].

The other instances in which opaque proxy discrimination can occur is when the proxy stands in for a true causative variable which, although understood, is difficult to quantify. Taking the example of online advertising and targeted marketing, user engagement on social media platforms is a known causative variable for effective advertising. However, measuring engagement precisely can be challenging due to its multifaceted nature. Although these metrics may not entirely represent the depth and quality of engagement, in this scenario, the quantity of likes, shares, or comments may act as a proxy for user engagement. Consequently, if an AI system solely relies on these proxies without considering alternative and presently quantifiable variables, it might inadvertently lead to opaque proxy discrimination in advertising practices [3].

*Indirect proxy discrimination* refers to instances in which AI models perform discriminatory actions not only based on direct connections but also via indirect correlations using prohibited variables. This type of discrimination occurs when a variable is used to predict an outcome not due to its inherent nature but rather because it serves as a proxy for another assessable variable despite not being in the AI's training data. An illustration of this is in loan approvals, where an AI model assesses creditworthiness based on the length of an individual's employment. Hence, this score can serve as a proxy for financial stability. In the realm of indirect proxy discrimination, the AI's predictions would not be based on the length of employment but rather on its relationship with more

[5] A situation where individuals may refrain from engaging in activities that are socially beneficial due to the perceived or actual negative consequences imposed by certain policies.

quantifiable and accurate variables, such as income stability. The suspect variable in these cases is not predictive on its own; rather, it offers a means by which to predict the probability of a true causative factor that is both potentially available and quantifiable [3].

The analysis of different types of proxy discrimination, whether causal, opaque, or indirect is a useful exercise. However, this predictability aspect of variables becomes increasingly complex in the field of big data. These algorithms rarely have a singular explanation for their predictions, thus the process of determining the probability of proxy discrimination in an AI system is quite challenging [3], [34].

Frequently, the cause of an outcome is a combination of multiple variables. For example, sex and mortality have a correlation, with women generally having a higher life expectancy than men. In this case, sex is a prohibited variable to base predictions on; however, algorithms may resort to facially neutral proxies, which may in fact act as indirect indicators of sex. These facially neutral proxies could include patterns of behaviour, preferences, or other observable traits that, while not explicitly related to sex, are correlated with it [3], [35].

With respect to predicting life expectancy, an AI proxying sex through social media information such as likes or names can include all three different types of proxy discrimination simultaneously. In terms of causal discrimination, this may occur since biological sex directly affects some elements of life expectancy. For opaque discrimination, sex may be proxied for unknown or immeasurable variables. Finally, indirect discrimination may occur for neutral variables such as smoking habits [3].

Various forms of proxy discrimination often coexist within the same AI model, with suspect variables serving as proxies for facially neutral variables, which in turn contributes to predicting the desired outcomes. Alternatively, an AI may cascade from one suspect variable to another, creating a chain of proxies for facially neutral variables linked to the target variable. Despite identifying AI-induced proxy discrimination as theoretically possible, a practical implementation of this proves to be challenging. This issue is especially prevalent when evaluating the social implications of AI-driven proxy discrimination, as discussed in the subsequent section [3].

## IV. Algorithmic Fairness: Concepts and Challenges

An algorithm is formally characterised as a predefined set of steps provided to a computer, constituting a sequence of instructions aimed at performing specific computational tasks or mapping an input domain to an output domain [36]. In the realm of machine learning models, training algorithms are instrumental in the creation of statistical models. This is done primarily by minimising the error between predicted labels and the actual outcomes corresponding to a set of input features.

### A. What is Algorithmic Fairness

Algorithmic fairness can be defined as the process of reducing the biases within machine learning models that could hurt or discriminate against groups of people. The main objective is to prevent unjust results, especially when it comes to protected characteristics like race, gender, and religion. Striving for fairness requires a deep understanding of any biases that might be present in algorithms, as well as being aware that these biases can unintentionally stem from the data used during training. Therefore, an important part of achieving fairness in machine learning is recognising where these biases originate from and highlighting how training data can affect the behaviour of AI models [37].

Machine learning engineers and data scientists are actively addressing biases in algorithms to promote fairness in decision-making. This involves not only technical improvements but also ethical and societal considerations. Fairness-focused techniques, such as the re-evaluation of feature importance and model adjustments, are employed to actively encourage inclusivity and prevent the continuation of unfairness. Achieving algorithmic fairness and creating a fair machine learning environment requires a mix of technical knowledge, ethical thinking, and awareness of societal needs [37].

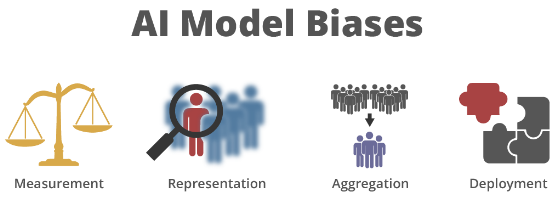


Fig. 5. Forms of AI model biases.

### B. Bias Definition and the Types of biases

In broad terms, bias encompasses the warping of an idea or concept that comes from disproportionate emphasis placed on supporting or opposing factors. These biases may appear inherently or be acquired through life experience (or, in the case of a learning model, through the training process), manifesting as either voluntary or involuntary inclinations. Acknowledging biases is essential, as they can lead to misinterpretations and conceal the true context of a situation.

In the field of machine learning, bias functions in a comparable manner, typically denoting either an imbalance within the dataset where certain features are assigned extra significance or a flaw in the model itself. This may lead to a skewed result, potentially perpetuating unfair outcomes. Recognising and mitigating bias in AI systems is essential to ensuring unbiased decision-making, as unchecked biases can have greater ramifications in various applications, from automated decision support to predictive analysis.

### B.1. Types of Biases

As AI models become increasingly integrated into our society, the implications of biased models are becoming more profound and far-reaching, and the potential consequences of biases in these models can impact various aspects of our daily lives. As illustrated in Figure 5, examples of biases within AI models come in many forms, as outlined in [38]:

### *1. Measurement Bias*

*Measurement Bias* arises when data accuracy differs among represented groups, often due to variations in proxy variables. Vigilant consideration and validation of these variables, particularly proxies, are crucial for ensuring consistent and reliable data representation across diverse groups.

An example of measurement bias is when survey results are collected from an urban area. Because the results were gathered from a limited geographic area, groups from more rural areas would be underrepresented, leading to unfairness [39].

An example of measurement bias is when a hospital uses a model to predict high-risk patients using factors like past diagnoses and medications, with healthcare costs as a proxy for risk, but under-identifies eligible black patients. This bias stems from the cost-risk relationship across races and is influenced by factors like barriers to care in healthcare, which result in lower average medical costs for black patients compared to patients of other races.

### *2. Representation Bias*

*Representation Bias* arises when the training data inadequately reflects the complexity of the real-world scenario, often due to insufficient variation between different groups or issues with the collection method. Addressing this bias is crucial for improving the model's ability to generalise effectively .

An example of this is in a medical dataset. If there is an underrepresentation of women, data on them becomes more sparse, and could lead to nuances between men and women to be lost. This could lead to a less accurate diagnoses for female patients compared to male patients [39].

### *3. Aggregation Bias*

*Aggregation Bias* arises when predictions for individual groups rely on characteristics of the entire population, potentially leading to an overly generic model that overlooks unique features. Addressing this bias is essential for a more accurate representation of diverse subgroups in model predictions.

An example of this is when a certain group of people has higher rates of diabetes and related complications than those of other groups. If an AI is built to diagnose diabetes, it is important to make the system sensitive to these ethnic differences by either including ethnicity as a feature or building separate models tailored for each group [38].

### *4. Deployment Bias*

*Deployment Bias* arises when a model is used for a purpose it was not initially designed for. Models trained for specific tasks may not perform well when applied to different tasks. Aligning the model's design with its intended use is crucial to minimise deployment bias and enhance performance across various applications.

An example of this is when the criminal justice system uses tools to predict the likelihood of a convicted individual to re-offend. These predictions are not intended for judges to consider when determining appropriate punishments during sentencing.

## *C. How to measure Algorithmic Fairness*

The idea of fairness is a subjective concept with many different definitions. This is a challenge prevalent in various disciplines, particularly in algorithmic decision-making systems. The absence of a consensus on what fairness is complicates the measurement of algorithmic fairness, as different interpretations lead to diverse evaluation methods. As a result, the development of precise and meaningful measurement techniques is a complex task.

Navigating this challenge requires a careful selection of measurement methods that align with specific interpretations of fairness. The following section will explore some of the most prominent measurement methods for assessing algorithmic fairness, offering distinct perspectives on this matter.

### *1. Disparate Impact*

*Disparate Impact* compares the proportion of individuals receiving positive outcomes between an unprivileged[6] group and a privileged[7] group. A high ratio between the positive prediction rates within these groups ensures similar positive prediction rates across groups, improving fairness [40].

Formally, disparate impact is calculated as follows:

$$\frac{P[\hat{Y}=1|S\neq 1]}{P[\hat{Y}=1|S=1]} \geq 1-\varepsilon \tag{7}$$

where S is the protected attribute (such as race, gender or religion), $S = 1$ refers to the privileged group, and $S \neq 1$ refers to the underprivileged group, and $\hat{Y} = 1$ indicates that the prediction is positive.

A higher value of this measure indicates more equitable rates, signifying increased fairness.

### *2. Demographic Parity*

*Demographic Parity*, also referred to as statistical parity, is similar to the previous metric, differing in its focus on the absolute difference rather than the ratio. Formally, it measures the contrast in outcomes between privileged and underprivileged groups by examining the absolute difference in positive predictions. This difference-based approach provides an alternative approach to assessing fairness and the disparities within algorithmic decision-making [41].

Formally, demographic parity is calculated as such:

$$|P[\hat{Y}=1|S=1]-P[\hat{Y}=1|S\neq 1]| \leq \varepsilon \tag{8}$$

A lower value of demographic parity signifies greater uniformity in acceptance rates, reflecting an enhanced level of fairness.

[6] A group of individuals that may obtain some unfair disadvantage.

[7] A group of individuals that may obtain some unfair advantage.

### 3. Equalised Odds

*Equalised Odds* addresses the limitations associated with measures like disparate impact and demographic parity. It calculates the disparities by evaluating the differences between the false positive rates and the true positive rates across the two groups. The equalised odds method offers a more nuanced perspective on fairness in comparison to traditional measures [42].

Formally, equalised odds is calculated as such:

$$|P[\hat{Y}=1|S=1,Y=0]-P[\hat{Y}=1|S\neq 1,Y=0]|\leq \varepsilon \quad (9)$$

$$|P[\hat{Y}=1|S=1,Y=1]-P[\hat{Y}=1|S\neq 1,Y=1]|\leq \varepsilon \quad (10)$$

where $Y$ indicates whether the ground truth prediction is positive or negative.

Reduced disparities between groups suggest a higher degree of fairness. It is crucial to note that the equalised odds method relies on the actual ground truth $Y$, assuming that the base rates of the two groups are representative and free from bias in their acquisition.

### 4. Equal Opportunity

Consistency in true positive rates across diverse groups is a fundamental requirement for achieving algorithmic fairness, as well as ensuring that individuals which attain positive outcomes consistently receive positive predictions from the model. This method extends beyond parity to uphold the principle of equal opportunity across demographic categories [42].

$$|P[\hat{Y}=1|S\neq 1,Y=1]-P[\hat{Y}=1|S=1,Y=1]|\leq \varepsilon \quad (11)$$

Highlighting the equality in one of the error types, such as true positives, will result in an increase in the difference between this error type and others, and can be problematic when groups have different error type base rates.

Combining and making use of all these metrics can facilitate the algorithmic decision-making evaluation process and offer a more holistic approach to achieving fairness in machine learning models.

## D. Causes of Algorithmic Unfairness

The causes of algorithmic unfairness are widespread and varied, and can stem from a multitude of different sources. The main sources of unfairness or bias in machine learning algorithms as outlined in [43] and [44] are presented below.

### 1. Selection Bias

*Selection Bias* occurs when the process of collecting data results in a misrepresentation of the true population it is attempting to model. This can occur if data collection is carried out in such a way that it causes a subset of the population to be less inclined to participate. This make take place in politics, where richer individuals are over-represented as they are more likely to participate in political events, result in misrepresented vote predictions.

### 2. Systematic Errors

*Systematic Errors* occur when the method of data collection in itself has some kind of flaw that could bias the data. An example of this is a faulty device used to measure some feature for the dataset. If the device is faulty, the data captured will not be the same as the true result.

### 3. Survey Bias

*Survey Bias* may occur when the method of data collection requires individuals to submit their own responses. The data can be skewed due to individuals' desire to appear socially acceptable, or due to the mistaken recollection of how past events took place.

### 4. Observer Bias

*Observer Bias* occurs when information is interpreted or collected in a way to favour a specific hypothesis or assumption rather than true reality. The interpretation of results could be skewed to fit with an individual's anticipated result, or the questions to gather data could be tailored, either intentionally or subconsciously, towards a specific assumed result.

### 5. Algorithmic Bias

*Algorithmic Bias* occurs due to the training algorithm itself rather than the data. It can occur for a variety of reasons, such as the incorrect application of a certain algorithm, a wrong approach to the problem in general, or an unsuited data format.

## E. Mitigating Algorithmic Unfairness

Mitigating algorithmic unfairness involves implementing strategies and practices throughout the entire machine learning model life cycle. In [45], a few mitigation strategies are proposed, shown below:

### 1. Correct Framing of the problem

When designing a prediction model, careful consideration of the research question, populations, predictors, and endpoints is essential. To deal with this, factors like available data, problem complexity, patient demographics, and community involvement must be considered fully, ensuring a comprehensive and unbiased approach.

### 2. Bias Detection and Evaluation

A rigorous bias detection and evaluation framework should be instituted, incorporating various fairness metrics to quantitatively assess disparate impacts. Utilising specialised tools and techniques for bias detection enhances the model's interpretability, such as ROC-AOC Analysis, cross-validation analysis and sampling techniques.

### 3. Explainability and Transparency

Emphasising the interpretability and transparency of the model's decision-making processes through explainable AI techniques fosters accountability. Rendering decision-making processes interpretable allows for external scrutiny by entities external to the team of model developers.

This research advocates for a comprehensive and iterative approach to addressing algorithmic unfairness grounded in ethical principles and a commitment to equity throughout the machine learning model's life cycle.

## V. Ethical Discussion on Algorithmic Decision-Making

As society continues to put more faith in the decisions taken by AI algorithms, the responsibility for developers to design ethical and transparent AI, as well as for individuals to be well informed on AI, is becoming increasingly paramount. The explainability of these algorithms is also necessary on the user side, so that individuals of all levels of technical ability and understanding may scrutinise and be aware of the algorithms that are taking potentially critical decisions that can affect their livelihood.

### A. Degrees of Impact/Risk

The degree to which the decision-making process of an AI model may be deemed unethical is typically reflected in the degree of impact it can have on the individuals that the decision concerns, i.e., what level of risk is introduced by automating this decision-making process. This notion is prevalent in the recently proposed EU AI Act [46], which aims to classify AI systems into different categories based on the risk they pose to society, as illustrated in Figure 6.

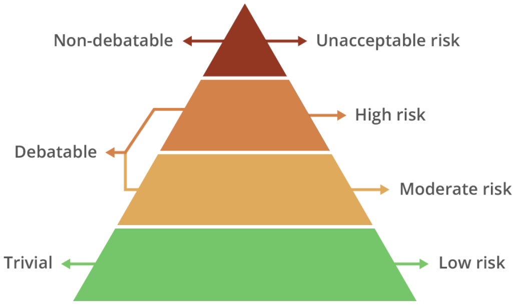


Fig. 6. Levels of AI model risk as defined by the EU AI Act and corresponding debatability.

Low-impact decisions are usually trivial and have little to no effect on users. Examples of these decisions include choosing what processing filter to put on an image and how to tweak a user interface's structure or colours. These types of decisions are artistic in nature or taken for the sake of convenience and are not subject to ethical discussion given their subjective nature and lack of impact on individuals with regards to the necessities for their livelihood, e.g., finance, health, and mental well-being.

AI models which take moderate-to-high-risk decisions may have social implications and a direct or indirect impact on individuals' rights [47], and thus are subject to ethical discussion. Examples of these decisions include machine learning model feature selection, medical diagnoses, self-driving vehicles, facial recognition, insurance provision, and criminal sentencing. These decisions have the potential to:

- Impact individuals' financial situations and job prospects [48].
- Impact how much priority they are given over other patients given how severe their health condition is deemed by the AI [49].
- Pose a direct life-threatening hazard, e.g. in the case of self-driving cars which choose to prioritise the well-being of the passengers over that of pedestrians [50].

### B. Algorithmic Fairness in a Biased Society

By definition, as provided in Section IV, it is impossible for an algorithm to be biased in itself as it is a predetermined sequence of steps that will always return the same output for a given input, assuming no outside influence. The statistical model generated as a result of a training algorithm, on the other hand, may be biased, either as a result of intentional or unintentional implementation choices made by a developer, or as a result of bias within the dataset used during training. If an AI model aims to perfectly represent a dataset built in a biased manner, then the choices and mappings made by that AI will also *seem* biased as it would have been programmed to make predictions as such – "garbage in, garbage out" [51].

#### B.1. Bias introduced by developers

Machine learning model developers are responsible for making a number of technical choices when it comes to training AI models, which affect the trajectory of the model:

#### 1. Feature Selection

Developers are responsible for choosing the most relevant features (input variables) [52] to train a model to extract a corresponding label (output). Features may be either overlooked (e.g. disregarding family medical history when predicting diagnoses) causing the model to lack critical information, or unnecessarily included such that they skew the model's mapping towards unjust discrimination (e.g. considering race or gender to determine the qualification of an individual for a job position, or information such as their zip code which may be indirectly indicative of their race due to geographic bias) [53].

#### 2. Algorithm Selection

There are multiple classes of machine learning algorithms, each capable of tackling specific tasks, including supervised learning[8], unsupervised learning[9] and neural networks[10]. Developers must choose algorithms according to the task at hand appropriately such that they are not excessively complex and prone to overfitting, particularly on biased data (e.g. training a deep neural network when linear regression would suffice, the neural network is more capable of learning the bias in the data [54]).

[8] Training a model on a set of features and their known label to make future predictions given only new features

[9] Pattern recognition in unlabelled data through grouping, clustering

[10] Fitting a network of connected nodes and weights to data by minimising the error between its predictions and the ground truth using back-propagation

*3. Hyperparameter Tuning*

Hyperparameter tuning concerns the tweaking of the predetermined parameters of the learning algorithm (not of the model) to maximise the performance of the generated model and optimise learning time. When compounded, the influences of each hyperparameter have the potential to exacerbate bias in the model's predictions [55]. Examples of hyperparameters and their potential to introduce bias and affect minorities include binary classification threshold (if set incorrectly, one class may be preferred over another), number of clusters in K-Means clustering (individuals may be incorrectly assigned to a specific group), dropout rate (if set too low, neural networks are more likely to learn the nuances and bias of the training data) and batch size (if set too high, neural networks are exposed to less noise and are less likely to generalise [56], potentially learning the bias in the data).

*4. Data Augmentation*

Data augmentation is used to artificially increase the quantity of data in a training set by generating new modified versions of existing data by applying certain transformations (e.g. in the case of computer vision, adjusting brightness, saturation, applying rotations, cutout, etc.). If these transformations are applied in an unbalanced manner (such as to a misrepresentative subset of the data consisting primarily of individuals of a specific race), a class imbalance will be created and if not addressed (e.g. through undersampling), models trained on this data may be skewed towards predicting this newly inflated class, underrepresenting other classes in the process [57].

*5. Evaluation Metrics*

Evaluation metrics gauge the overall performance of machine learning models. It is paramount that the appropriate metrics are used depending on the context in which the AI models are used, as individuals may otherwise be negatively impacted by a model which is wrongly reported to be performing well [58]. Examples of incorrect evaluation metric usage include using a macro average to evaluate multi-class classification when there is an imbalance between the quantity of classes in the training data [59], and using accuracy as the only evaluation metric when making disease diagnoses. Ideally recall and precision are also considered and recall is given more importance to minimise the number of false negatives which would result in infected individuals not being given the necessary treatment.

Improper hyperparameter choices, while not necessarily having drastic effects in their own right, have the potential to compound and result in a heavily biased model. Ideally, AI models are to be developed by a diverse team of developers to minimise the potential for individual bias.

### B.2. Bias introduced during dataset curation

As discussed in Section IV, bias in datasets can arise in a multitude of ways. Even if data is collected in a seemingly equal manner, bias may still be introduced by the necessity of participants' subjective responses, unintentional sampling or confirmation bias, and response bias, among others. Given the aim of creating an AI model that is as objective as possible, responsibility for addressing the problem of dataset bias can be split three-fold: collection, curation and detection.

*1. Collection*

The individuals, organisations or institutions gathering data are tasked with collecting data in a manner that is free from systemic bias. They are responsible for ensuring diverse sampling of data and preempting predictable sources of bias[11] and adjusting the sampling approach as necessary to mitigate them.

Participants, if directly involved and fully consenting, are also responsible for providing accurate and truthful responses so that the dataset is representative of the current, real-world societal situation.

*2. Curation*

Individuals tasked with the curation[12] of datasets through collected data should be as objective as possible and are responsible for assuring the quality of said data, including looking out for missing values, outliers, and duplicate entries. Statistical analysis should also be carried out to detect imbalanced data and biases that may have been overlooked during collection [60].

*3. Detection*

Once a dataset has been implemented and used to train a machine learning model, it is the responsibility of data scientists and machine learning engineers to continuously monitor the predictions made by this model and look out for bias in the model's decisions [60]. If bias is found, it should be addressed immediately by tweaking the dataset and retraining the model such that the bias is mitigated, all while ensuring that use of the biased model is halted to prevent further discrimination and potential harm to individuals.

### B.3. Biased AI Models vs. Subjective Human Judgement

The discourse regarding the ethics of AI model decision-making, therefore, converges towards one question. Would it be more beneficial for society to utilise AI models with the potential for bias which can be mitigated through technical interventions, or to rely on the judgement of humans assuming they are acting in good faith and as objectively as possible? Presented below is a comparison of both approaches when taking into account multiple considerations:

*1. Consistency*

AI models, unless programmed otherwise or utilising probabilities, provide deterministic, reproduceable mappings from inputs to outputs [61], [62]. Even if a model gives biased results due to the reasons presented in Subsection V-B, it will be consistently biased, and therefore this discrepancy can be detected and treated accordingly.

[11] e.g. over/under-representation, temporal bias, geographic bias

[12] Labelling, annotation, analysis

Human judgement has the potential for variability and inconsistency due to external factors such as mood, tiredness, and shifting beliefs.

### 2. *Bias*

AI models, as discussed in Subsection V-B, can be biased either as a result of learning from a biased dataset or due to developer error. Either case can be addressed through technical intervention by building a more representative and fair dataset or by tweaking and retraining the model to mitigate observed bias; therefore, the objectivity of AI models can be continuously optimised.

Humans are subjective by nature due to a variety of factors, including:

- Differences in perception and thought processes as a result of different biological structures and experiences
- Discrepancies in intended meaning and corresponding interpretation, relative beliefs based on culture and religion [63]
- Individual values and the subjectivity of language as a tool itself.

They can, and in many cases, are expected to, minimise the impact of their personal beliefs when carrying out objective judgement of a case/situation, though an element of subjectivity is unavoidable [64], [65].

### 3. *Emotional Intelligence*

Emotion understanding can play a critical role in making certain decisions, especially in legal contexts such as determining witness honesty during testimonies and distinguishing between premeditation and accidental actions carried out by criminals. These emotions are reflected in complex nuances in speech, facial expressions, tone and hesitation which are hard to express numerically in an objective manner as individuals' traits vary from person to person. Thus, it can be challenging for an AI model to accurately and reliably capture the intricacies of all human emotions [66].

Humans' complex neurological system [67] and exposure to social contexts from a young age [68] make them significantly more proficient at emotion recognition. In legal contexts, for example, the jury (in the case of trial by jury) provides power in numbers, combining the subjective interpretation of multiple jury members into a unanimous or majority verdict. Judges are also exposed to a wide array of character traits over time and develop a heightened ability to distinguish between emotions [69], as well as whether or not they are being portrayed genuinely.

AI models, given their consistency and predictability, as well as curators' ability to combat bias during dataset development and model training, prove to be good tools for aiming towards objective decision-making. Regardless, there are contexts where objectivity does not necessarily imply fair treatment, and in these cases, human intervention still proves to be necessary. Ultimately, the collaborative use of AI models and human judgement is a balanced approach. AI contributes technical efficiency and objectivity, while humans provide the nuanced, context-specific understanding necessary for fair and ethical decision-making in diverse and dynamic situations. Striking this balance ensures that technology augments human capabilities without compromising fundamental principles of justice and fairness.

## VI. Conclusion and Future Directions

### A. *Proxies*

Proxies, while serving as convenient substitutes for sensitive attributes, can inadvertently perpetuate biases and contribute to unfair decision-making. The recognition of proxies as a potential source of bias necessitates a proactive approach to algorithmic design, demanding careful consideration of the features used during model training.

Addressing the challenges associated with proxies requires a multifaceted strategy that encompasses robust data collection practices, algorithmic transparency, and continuous monitoring. Developers must exercise vigilance in selecting features to avoid unintentional correlations with sensitive attributes. Additionally, ongoing efforts in the research community and industry collaborations are crucial for identifying and mitigating the impact of proxies on algorithmic fairness.

As the field of AI advances, future directions in addressing proxies should involve the development of standardised guidelines, ethical frameworks, and tool sets that empower practitioners to navigate the complexities of proxy-related biases effectively. Fostering a culture of awareness and responsibility within the AI community will contribute to the evolution of fair and unbiased AI systems, aligning with the broader goal of creating technology that benefits all of society.

### B. *Algorithmic Fairness and the Ethics of Algorithmic Decision-Making*

The presented discussion of algorithmic fairness and ethical considerations in AI decision-making underlines their importance in the shaping of responsible and transparent AI practices. This discussion has provided insights into the formal characterisation of algorithms, methods for measuring and assessing algorithmic fairness and different causes of algorithmic unfairness. A variety of strategies for the mitigation of algorithmic unfairness are proposed.

Ethical discussions in algorithmic decision-making are driven by the varying degrees of impact and risk that AI models posed to individuals. The recent EU AI Act's categorisation of AI systems based on potential risk highlights the growing necessity for responsible AI deployment. Bias introduced by developers and during dataset curation outlines the significance of informed choices, responsible data practices, and the continuous monitoring of AI systems.

While AI models offer consistency and potential for optimisation, they may encounter challenges in understanding complex human emotions. Human judgement, enriched by emotional intelligence, remains indispensable in certain contexts, advocating for a balanced approach.

Achieving algorithmic fairness and navigating ethical considerations in AI requires a balanced and continuous approach. It requires a commitment to transparency, ongoing scrutiny, and a collaborative effort between technological advancements and human understanding. The ethical deployment of AI models emerges as a shared societal responsibility, demanding interdisciplinary collaboration and an ongoing commitment to ethical practices in AI integration.

## References

[1] L. Floridi, "A proxy culture," http://dx.doi.org/10.2139/ssrn.3839315, October 21 2015.

[2] G. Choueiry, "Quantifying health," Jan 2020.

[3] A. Prince and D. Schwarcz, "Proxy discrimination in the age of artificial intelligence and big data," *Iowa Law Review*, vol. 105, p. 1257, 2020, 62 Pages Posted: 17 Apr 2019 Last revised: 9 Apr 2020.

[4] M. C. Tschantz, "What is proxy discrimination?" in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1993–2003.

[5] T. Erickson and W. A. Kellogg, "Social proxies," *The Encyclopaedia of Human Computer Interaction, Berkshire: Berkshire Publishing Group LLC*, 2004.

[6] J. Pickering, "Cognitive engagement: A more reliable proxy for learning?" *Medical Science Educator*, vol. 27, no. 4, pp. 821–823, 2017.

[7] W. Grove, T. Wasserman, and A. Grodner, "Choosing a proxy for academic aptitude," *Journal of Economic Education*, vol. 37, pp. 131–147, 02 2006.

[8] T. Brown-Nagin, "Rethinking diversity and proxies for economic disadvantage: A first generation students' project," *Chicago Legal Forum*, forthcoming 2015, this article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-useOAP.

[9] R. D. Waikar, S. S. Kalagnanam, and I. M. Findlay, "Financial proxies for social return on investment analyses in saskatchewan," Saskatoon, SK, Canada, 2013.

[10] K. Dynan and L. Sheiner, "Gdp as a measure of economic well-being," Hutchins Center, The Brookings Institution, The Brookings Institution, Tech. Rep. 43, August 2018.

[11] R. Boarini, Åsa Johansson, and M. M. damp;apos;Ercole, "Alternative measures of well-being," no. 33, 2006.

[12] D. S. Morris, D. Schwarcz, and J. C. Teitelbaum, "Do credit-based insurance scores proxy for income in predicting auto claim risk?" *J. Empirical Legal Stud.*, vol. 14, no. 2, pp. 397–423, 2017, 33 Pages Posted: 3 Nov 2015 Last revised: 25 Jan 2018.

[13] E. Volokh, "Diversity, race as proxy, and religion as proxy," 1997, 17 Pages Posted: 1 Feb 1997.

[14] M. W. Foster and R. R. Sharp, "Race, ethnicity, and genomics: Social classifications as proxies of biological heterogeneity," *Cold Spring Harbor Laboratory Press*, 2002, department of Anthropology, University of Oklahoma, Norman, Oklahoma 73019, USA; Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma 73104, USA; National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709-2233. Abstract available at http://www.genome.org/cgi/doi/10.1101/gr.99202. Corresponding author: Morris.W.Foster-1@ou.edu; FAX (405) 325-7386.

[15] D. Carr and D. Khodyakov, "Health care proxies: Whom do young old adults choose and why?" *Journal of Health and Social Behavior*, vol. 48, no. 2, pp. 180–194, 2007, pMID: 17583273.

[16] C. M. Doak, D. J. Hoffman, S. A. Norris, M. Campos Ponce, K. Polman, and P. L. Griffiths, "Is body mass index an appropriate proxy for body fat in children?" *Global Food Security*, vol. 2, no. 2, pp. 65–71, 2013.

[17] A. R. Heid, L. R. Bangerter, K. M. Abbott, and K. V. Haitsma, "Do family proxies get it right? concordance in reports of nursing home residents' everyday preferences," *Journal of Applied Gerontology*, vol. 36, no. 6, pp. 667–691, 2017, pMID: 25926658.

[18] L. S. Williams, T. Bakas, E. Brizendine, L. Plue, W. Tu, H. Hendrie, and K. Kroenke, "How valid are family proxy assessments of stroke patients' health-related quality of life?" *Stroke*, vol. 37, no. 8, pp. 2081–2085, 2006.

[19] C. Izzo, Z. A. Doubleday, G. L. Grammer, K. L. Gilmore, H. K. Alleway, T. C. Barnes, M. C. F. Disspain, A. J. Giraldo, N. Mazloumi, and B. M. Gillanders, "Fish as proxies of ecological and environmental change," *Reviews in Fish Biology and Fisheries*, vol. 26, no. 3, pp. 265–286, 2016.

[20] Z. J. N. Steinmann, A. M. Schipper, M. Hauck, S. Giljum, G. Wernet, and M. A. J. Huijbregts, "Resource footprints are good proxies of environmental damage," *Environmental Science & Technology*, vol. 51, no. 11, pp. 6360–6366, 2017.

[21] R. Patalano and P. Roberts, *Climate Proxies*. John Wiley Sons, Ltd, 2021, pp. 1–5.

[22] A. Todorov and C. Kirchner, "Bias in proxies' reports of disability: Data from the national health interview survey on disability," *American Journal of Public Health*, vol. 90, no. 8, pp. 1248–1253, 2000.

[23] J. Reynolds and J. B. Wenger, "He said, she said: The gender wage gap according to self and proxy reports in the current population survey," *Social Science Research*, vol. 41, no. 2, pp. 392–411, 2012.

[24] D. Mulvin, *Proxies: The Cultural Work of Standing In*, 08 2021.

[25] J. Younis, H. Freitag, J. S. Ruthberg, J. P. Romanes, C. Nielsen, and N. Mehta, "Social media as an early proxy for social distancing indicated by the covid-19 reproduction number: Observational study," *JMIR Public Health Surveill*, vol. 6, no. 4, p. e21340, Oct 2020.

[26] C. Deng, W. Lin, X. Ye, Z. Li, Z. Zhang, and G. Xu, "Social media data as a proxy for hourly fine-scale electric power consumption estimation," *Environment and Planning A: Economy and Space*, vol. 50, no. 8, pp. 1553–1557, 2018.

[27] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo, "Predicting bounce rates in sponsored search advertisements," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 1325–1334.

[28] E. P. Goss and J. M. Phillips, "How information technology affects wages: Evidence using internet usage as a proxy for it skills," *Journal of Labor Research*, vol. 23, no. 3, pp. 463–474, September 1 2002.

[29] X. Rao, H. Zhao, and Q. Deng, "Artificial-neural-network (ann) based proxy model for performances forecast and inverse project design of water huff-n-puff technology," *Journal of Petroleum Science and Engineering*, vol. 195, p. 107851, 2020.

[30] R. DePaula, "A new era in human computer interaction: The challenges of technology as a social proxy," in *Proceedings of the Latin American Conference on Human-Computer Interaction*, ser. CLIHC '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 219–222.

[31] C. Webster, S. Taylor, C. Thomas, and J. Weller, "Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations," *BJA Educ*, vol. 22, no. 4, pp. 131–137, Apr 2022.

[32] Target corporation to pay $2.8 million to resolve EEOC discrimination finding. U.S. EEOC. No date.

[33] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, 08 2021.

[34] B. Davies and T. Douglas, "Learning to discriminate: The perfect proxy problem in artificially intelligent criminal sentencing," in *Sentencing and Artificial Intelligence*, J. Ryberg and J. V. Roberts, Eds. Oxford University Press, 2022.

[35] L. Alexander and K. Cole, "Discrimination by proxy," *Constitutional commentary*, vol. 14, no. 3, pp. 453–463, 1997.

[36] R. K. Hill, "What an algorithm is," *Philosophy & Technology*, vol. 29, pp. 35–59, 2016.

[37] D. Pessach and E. Shmueli, "Algorithmic fairness," 2020.

[38] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[39] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," 2023.

[40] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 259–268.

[41] L. Rosenblatt and R. T. Witter, "Counterfactual fairness is basically demographic parity," 2023.

[42] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016.

[43] F. Martínez-Plumed, C. Ferri, D. Nieves, and J. Hernández-Orallo, "Fairness and missing values," 2019.

[44] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," 2018.

[45] L. H. Nazer, R. Zatarah, S. Waldrip, J. X. C. Ke, M. Moukheiber, A. K. Khanna, R. S. Hicklen, L. Moukheiber, D. Moukheiber, H. Ma, and P. Mathur, "Bias in artificial intelligence algorithms and recommendations for mitigation," *PLOS Digital Health*, vol. 2, no. 6, pp. 1–14, 06 2023.

[46] "EU AI Act: first regulation on artificial intelligence | News | European Parliament," aug 2023.

[47] R. Rodrigues, "Legal and human rights issues of ai: Gaps, challenges and vulnerabilities," *Journal of Responsible Technology*, vol. 4, p. 100005, 2020.

[48] M. A. Malek, "Criminal courts' artificial intelligence: the way it reinforces bias and discrimination," *AI and Ethics*, vol. 2, no. 1, pp. 233–245, 2022.

[49] S. Hoffman and A. Podgurski, "Artificial intelligence and discrimination in health care," *Yale J. Health Pol'y L. & Ethics*, vol. 19, p. 1, 2019.

[50] S. Gless, E. Silverman, and T. Weigend, "If robots cause harm, who is to blame? self-driving cars and criminal liability," *New Criminal Law Review*, vol. 19, no. 3, pp. 412–436, 2016.

[51] G. M. Johnson, "Algorithmic bias: on the implicit biases of social technology," *Synthese*, vol. 198, no. 10, pp. 9941–9961, 2021.

[52] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[53] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[54] L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1827–1858.

[55] A. F. Cruz, P. Saleiro, C. Belém, C. Soares, and P. Bizarro, "Promoting fairness through hyperparameter optimization," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1036–1041.

[56] H. Yong, J. Huang, D. Meng, X. Hua, and L. Zhang, "Momentum batch normalization for deep learning with small batch size," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 224–240.

[57] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 18–19.

[58] E. Amigó, D. Spina, and J. Carrillo-de Albornoz, "An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 625–634.

[59] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[60] C. J. Pannucci and E. G. Wilkins, "Identifying and avoiding bias in research," *Plastic and reconstructive surgery*, vol. 126, no. 2, p. 619, 2010.

[61] I. Icke and J. C. Bongard, "Improving genetic programming based symbolic regression using deterministic machine learning," in *2013 IEEE Congress on Evolutionary Computation*. IEEE, 2013, pp. 1763–1770.

[62] D. Poole and A. E. Raftery, "Inference for deterministic simulation models: the bayesian melding approach," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1244–1255, 2000.

[63] R. Fernández and A. Fogli, "Culture: An empirical investigation of beliefs, work, and fertility," *American economic journal: Macroeconomics*, vol. 1, no. 1, pp. 146–177, 2009.

[64] B. G. Silverman, "Modeling and critiquing the confirmation bias in human reasoning," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 5, pp. 972–982, 1992.

[65] W. De Neys and J.-F. Bonnefon, "The 'whys' and 'whens' of individual differences in thinking biases," *Trends in cognitive sciences*, vol. 17, no. 4, pp. 172–178, 2013.

[66] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[67] E. T. Rolls, "Limbic systems for emotion and for memory, but no single limbic system," *cortex*, vol. 62, pp. 119–157, 2015.

[68] S. Denervaud, C. Mumenthaler, E. Gentaz, and D. Sander, "Emotion recognition development: Preliminary evidence for an effect of school pedagogical practices," *Learning and Instruction*, vol. 69, p. 101353, 2020.

[69] S. Porter and L. Ten Brinke, "Dangerous decisions: A theoretical framework for understanding how judges assess credibility in the courtroom," *Legal and Criminological Psychology*, vol. 14, no. 1, pp. 119–134, 2009.