Rolling in the deep of cognitive and AI biases

ATHENA VAKALI, Aristotle University, Greece

NICOLETA TANTALAKI, Aristotle University, Greece

Nowadays, we delegate many of our decisions to Artificial Intelligence (AI) that acts either in solo or as a human companion in decisions made to support several sensitive domains, like healthcare, financial services and law enforcement. AI systems, even carefully designed to be fair, are heavily criticized for delivering misjudged and discriminated outcomes against individuals and groups. Numerous work on AI algorithmic fairness is devoted on Machine Learning pipelines which address biases and quantify fairness under a pure computational view. However, the continuous unfair and unjust AI outcomes, indicate that there is urgent need to understand AI as a sociotechnical system, inseparable from the conditions in which it is designed, developed and deployed. Although, the synergy of humans and machines seems imperative to make AI work, the significant impact of human and societal factors on AI bias is currently overlooked. We address this critical issue by following a radical new methodology under which human cognitive biases become core entities in our AI fairness overview. Inspired by the cognitive science definition and taxonomy of human heuristics, we identify how harmful human actions influence the overall AI lifecycle, and reveal human to AI biases hidden pathways. We introduce a new mapping, which justifies the human heuristics to AI biases reflections and we detect relevant fairness intensities and inter-dependencies. We envision that this approach will contribute in revisiting AI fairness under deeper human-centric case studies, revealing hidden biases cause and effects.

1 INTRODUCTION

AI technology has invaded everyday life and activities, disrupting societal, economical, and policy making norms. AI acts either in solo or as a human companion in decisions made in so many domains: to provide health insights, approve loans, detect potential criminals and terrorists, screen University admissions and job applications, allocate digital content and so on [15, 21, 34]. The AI systems' ability to acquire cognitive skills such as perception, reasoning, and decision-making makes AI a valuable technology [15, 41]. However, the hypothesis that AI's mathematical and statistical nature will be more neutral, objective, and rational than humans, is not evident yet! In so many real-world cases there are several problems, associated with AI algorithms, based on the fact that they are often proven to be biased [36, 43, 46].

Up to now, bias and fairness terms are perplexed, studied under varying hypotheses and addressed with ad-hoc definitions and divergent goals. For example, in statistics, bias is a systematic error in the estimation of parameters or variables [64]. In this context, the effectiveness of AI systems was based on this definition and assessed based on how well their outputs correspond to their inputs. Thus, AI to be fair, meant that it had to deliver high accuracy based on the training data chosen and effectively generalize to unseen data. Apparently, the full spectrum of risks posed by bias in AI cannot be encompassed sufficiently by such a statistical perspective [55]. AI as a human mirroring technology, reflects both the fair and unfair patterns in society, and the perspectives of its creators, whether these perspectives are fair or biased [66].

In pursuit of Fairness in Artificial Intelligence (FairAI), a lot of research efforts explored the importance of data in mitigating bias and creating equitable outcomes. Conceptualizing "fairness" necessitated examining various data characteristics, like demographic information, from a technical perspective [16, 37]. Computational methods, though, like data re-sampling or re-weighting prove to be not enough, are costly, and demand prerequisites

Authors' Contact Information: Athena Vakali, avakali@csd.auth.gr, Aristotle University, Thessaloniki, Greece; Nicoleta Tantalaki, tantalaki@csd.auth.gr, Aristotle University, Thessaloniki, Greece.

that are not feasible in real-world datasets [30]. Rich and emerging bibliography propose solutions that are, though, generally grounded on computational AI pipelines, presume specific bias types over the phases of data collection (pre-processing), modeling (in-processing), and deployment (post-processing) phases, propose technical methodologies as bias mitigation strategies, and choose among numerous fairness metrics of questionable efficiency that are challenged by decision makers' own bias [6, 16, 18, 28, 42, 61, 63]. Although such studies have always been crucial for bias mitigation, they cannot guarantee a bias-free model [56, 63]. Unfortunately, even when a system is carefully designed to be fair, AI biases remain and harm minority or unprivileged populations (based on gender; age; vulnerable; and other traits), making the need to understand AI systems as sociotechnical systems urgent.

Recently, due to the increasing deployment of AI applications in sensitive domains like health-care and law enforcement, practitioners and scholars underlined the complexity of managing AI bias and the necessity to include behavioral and social aspects in their examination [6, 13, 18, 41, 42, 51, 56, 59, 63]. The collaboration of both humans and AI is essential to make this technology work as intended but the significant impact of human factors on AI bias is currently overlooked [41, 55, 56, 59, 63]. Human-AI synergy settings, must carefully consider that AI decisions and outputs are influenced by human systematic errors in thinking, which are known as cognitive biases. Cognitive biases relate to the outcome of utilising heuristics while thinking [3, 44], are common to all human beings, and influenced by their own reality, that is based on cognitive limitations, motivational factors, and/or adaptation to natural environments [62]. Given the strong human tendency to many bias types that often operate unconsciously [50], it has been noticed that human biases are not easily formulated and promptly recognized. Thus, they often "creep" into each phase of the AI lifecycle with unpredictable and complicated ways [13, 63]. According to the literature, more than 180 cognitive biases have been identified over the last six decades of research on human judgment and decision-making in cognitive science and this list is continually evolving [60].

In this work, we explore the knowledge about bias and fairness, coming from the mature area of research in cognitive science, and we argue that to introduce an actual human-centric FairAI demands a deep study of human and AI biases traits, patterns, and dynamics. We aim to provide a stepping stone in revealing the cognitive heuristics that underlie many identified biases in AI. To do so, we follow a systematic methodology, which : (a) *s*ets human and AI bias correlation assumptions and proceeds with deep hypothesis testing ; (b) *u*ncovers the particular heuristics which lead to specific harmful actions at each phase of the AI lifecycle (pre-processing, in-processing, post-processing) [30, 42, 61, 63]; (c) *p*roposes a systematic mapping under an explorative analysis among human heuristics, harmful actions, and specific distorted and harmful biased outcomes. In such a systematic manner, specific AI biases are recognized as reflections of these heuristics and hidden intensities and inter-dependencies among them are revealed. Our work relies on the scientifically grounded heuristics, introduced and mentioned in Daniel Kahneman's prize-awarded research with other researchers [23, 35, 36, 62]. The AI biases (or computational biases) discussed in this study are some of the most relevant in the context of AI and the most studied in literature [18, 38, 42, 61, 63]. Overall, we aim to offer new insights and a new theoretical framework which will roll in the deep of biases and fairness origins and dynamics, towards introducing a new line of human-centric FairAI research. At such a research line, we will bridge the knowledge and the evidence coming from cognitive and AI sciences, recognizing that it's not possible to either quantify human biases or to humanize AI biases. Thus, we propose a solid methodology for systematically mapping the human and AI biases, in support to a novel FairAI research practice which can no longer be purely computational, and which must be adaptive and human-inclusive in an end-to-end manner.

2 HUMAN HEURISTICS AND BIASES BEHIND DECISION-MAKING

A recent NIST special publication about standards for bias and fairness, has positioned computational biases at the tip of a biases iceberg, emphasizing current research inefficiency to deeper study human and systemic (real world) biases [55]. It's now urgent to go deep in the bias iceberg, taking advantage of the long studied topics and the scientific evidence revealed by cognitive science about human cognition and biased opinion making. Identifying causes of cognitive biases, though, is too challenging. In the groundbreaking work of Tversky and Kahneman "Judgment under Uncertainty" back in 1974 researchers argued that a broad family of cognitive biases can be explainable in terms of three heuristics; *representativeness*, *availability*, and *anchoring and adjustment* [62] (also addressed in their collection of papers [12]). Later in 2002, in "Heuristics and Biases: The Psychology if Intuitive Judgement" [23], Thomas Gilovich, Dale Griffin, and Daniel Kahneman have pulled together a new collection of papers to revisit each of these three major general purpose heuristics, reinforcing their importance. The Nobel Prize in Economic Science awarded to Kahneman in 2002 is certainly one indicator of the tremendous impact of heuristics and biases research. Another major theme underlined in this collection ([23]) is the fact that intuition should also be explained considering the role of affect, mood, and emotion. The *affect* heuristic was introduced by Paul Slovic et al. [58] and later mentioned in more Kahneman's works [35, 36]. These heuristics are briefly summarized next, to inspire our new FairAI hupothesis testing base.

Representativeness is about estimating the probability of an event by comparing it to a known situation or a prototype that already exists in someone's mind. People overweight certain information because they believe it's representative of reality. This may happen because this piece of information is more recent, or has been emphasized more in the media, or has been presented in a certain way. It's a mental shortcut that helps people group the world into simple categories like X and Y, but when a new experience or data rises, they cannot consider that it could be an entirely new category Z. The representativeness heuristic leads to the bias that makes people believe that a stereotype is true. A prominenet example of bias introduced by representativeness is the *illusion of validity* that refers to the overconfidence that is produced when we have a good fit between the predicted outcome and the input information [23, 35, 62].

Availability is a type of mental shortcut that involves assessing the frequency or plausibility of something based on how easily examples are brought to mind, without making an effort to look for other, potentially more significant information. Instances of large classes are usually recalled better and faster than instances of less frequent classes, likely occurrences are easier to imagine than unlikely ones, and the associative connections between events are strengthened when the events usually co-occur. People use this heuristic for estimating the number of elements in a class, the likelihood of an event, or the frequency of co-occurrences, by how easily the relevant mental operations of retrieval, construction, or association can be performed. A prominent bias that belongs in this category is the *illusory correlation* that makes people conclude that variables are correlated because their pairings come to mind easily (e.g. because they are quick to grasp, or because they seem likely) [23, 35, 62].

Anchoring and adjustment occurs when people base their initial ideas, estimates or predictions on a piece of information and make changes driven by this starting point. The anchoring bias refers exactly to this tendency of people to rely heavily on a reference point that is called "anchor" (and can be completely irrelevant e.g. an arbitrary number) and subsequently adjust that information until a final decision. Often, these adjustments are inadequate and remain close to the original anchor [62]. Automation bias is a well-known bias in the context of AI that belongs in this category. It occurs when individuals rely heavily on automated systems or technology to make a decision [35, 36]. Individuals may sometimes gravitate toward confirming their initial views which are the anchor, making insufficient adjustments, selectively gathering information that reinforces their starting point, giving rise to a well studied cognitive bias, known as confirmation bias [22, 39]. This type of bias is capable of involving both information processing and emotions as discussed below [14]. It should also be noted

that overconfidence's cognitive causes are among others the anchoring and the confirmation bias, as people anchor in a value or an idea without making the necessary adjustments. Instead, a competent strategic thinker would look for disconfirming evidence and alternative explanations for those signals that are ignored by others [10, 52].

Affect involves making decisions based on emotions and feelings associated with the given options. Decision makers' fast reactions to decision options, rather than analytical reasoning, serve as input to decisions. The affect heuristic is a key example of Kahneman's fast and intuitive i.e. fast so called "System 1 thinking" [35], where people jump to a conclusion and bypass the process of gathering information, giving rise to a type of bias called *conclusion* bias or prejudgement. However, people may also engage in deliberate thinking (mobilizing the so called "System 2 thinking", i.e. the one that is slow, conscious and requires intentional effort [35]), and their emotional comfort of having consistent beliefs, drive their tendency to seek arguments and favor information that confirm preexisting beliefs, giving once again rise to confirmation bias. [35, 36].

3 COMPUTATIONAL BIAS TYPES IN THE AI LIFECYCLE

Cognitive biases (we also refer to them as "human biases") described above, may enter the engineering and modeling processes and are ubiquitous in the decision making processes across the whole AI lifecycle, and in the use of AI applications once deployed [55]. In the context of AI, the typical approach of employing FairAI research and practices involves exploring specific bias types (known as "computational bias") throughout the AI lifecycle [18, 38, 42, 61, 63]. The following seven computational bias types are distinguished as the most prominent in the context of AI and the most studied in literature:

- **Historical bias** results from practices that lead to a specific treatment towards certain social groups being advantaged or favored, while others may be disadvantaged or excluded. Biases in this category arise even if data is flawlessly sampled and can be present in the datasets used prior to the creation of the model. They are the result of the majority following existing rules or norms. Racial biases, gender biases, biases towards people with disabilities belong in this category [55, 62, 63].
- **Representation bias** occurs when building datasets to sample from a population. A non-representative dataset lacks the diversity of the population. The sample might be small, wrongly sampled, with a distribution that differs from the true underlying distribution of the relevant population, and neglect underrepresented groups. Even with perfect a sampling technique, a model can have bad performance for the underrepresented groups dew to possible skewness of the underlying distribution. A model trained with fewer data points regarding a subgroup will not generalize well for it [42, 57, 61, 63]. Sampling bias mentioned in statistics belongs in this category and arises when the chosen sample represents a skewed subset of the target population [38, 61].
- **Measurement bias** arises from the choices made when choosing and computing particular features and labels. Usually, poor reflections are used as proxies of complex constructs. The method and accuracy of measurement may also vary across groups, found in the relevant population [38, 61]. As an example, using historical school performance as a proxy to predict the abilities of students, who could not take proper exams during COVID-19, introduced measurement bias that unfairly affected them.
- **Aggregation bias** arises when false conclusions are drawn about individuals from observing the entire population. Such biases come up even when different subgroups are represented equally in the training data but are inappropriately combined. Any general assumptions about subgroups should be avoided [42, 61]. For example, models for diagnosing and monitoring diabetes have relied on Hemoglobin A1c (HbA1c) levels for predictions. However, a recent study revealed that HbA1c levels vary in complex ways across different ethnicities [7], making it likely that a single model for all populations would display aggregation bias.

- **Algorithmic bias** occurs due to inappropriate technical considerations taken while building a model and have consequences related to fairness. Such choices regard aspects ranging from dataset split to different sets (e.g. training set, test set) to model's architecture and choices regarding optimization functions and hyper-parameters. Several mathematical abstractions are needed to develop a model and are made without prejudice or discriminatory intent [63, 64]. However, they disproportionately impact error rates on the underrepresented groups of the dataset. The key reason that these modeling choices lead to bias has to do with the fact that fairness often coincide with how underrepresented and protected features are treated by the model [30].
- **Evaluation bias** occurs when the benchmark population that is used during model evaluation, is not representative of the target population and/or inappropriate performance metrics are used [63]. Let's consider a smile detector model trained on a dataset lacking adequate representation of Asian individuals. If the testing benchmark is similarly unbalanced, mirroring the training set, the bias against Asians will go unnoticed [38]
- **Deployment bias** arises when the problem the model is designed to solve is different from the way it is actually utilized. This often occurs when a system is built and evaluated as if it were fully autonomous as human reviewers may overestimate its credibility leading to automation bias [61].

4 SYSTEMATIC MAPPING BETWEEN HUMAN HEURISTICS AND COMPUTATIONAL BIASES

In our methodology, we explore specific human actions as sources of biases which are then spotted at the pre-processing, in-processing, and post-processing phases of the AI lifecycle [45]. The selected harmful actions resulted from our exploratory analysis of the most popular computational biases (Section 2) commonly highlighted in several recent surveys [18, 38, 42, 61, 63]. In tandem with the Information Commissioner's Office (ICO) Annex regarding "Fairness in the AI lifecycle" [45], that sets out potential sources of bias, we harvest the necessary information around the decisions that may lead to unfair outcomes in the context of AI. Building on the grounded evidence that humans may take wrong or sub-optimal decisions, leading to problematic *actions*, based on the necessary but (often) biased four *heuristics*, we view computational biases as reflections of these human heuristics (Section 2). Then the explored human, i.e. cognitive biases are *reflected* to the prominent computational biases (Section 3), and an insightful mapping among them is provided. Our approach offers a solid scientific evidence coming from our rolling in the deep interdisciplinary study, which bridges cognitive and AI scientific knowledge, and reveals hidden human and AI ties which have been largely overlooked.

4.1 Pre-processing

In any typical AI pipeline, data must first be collected and a target population is identified, along with a set of features and labels chosen. Surely, the large scale of data, prohibits the inclusion of the entire data population, thus certain groups may be excluded or underrepresented, leading to inaccurate results that are harmful when making important decisions.

Harmful action 1: Use of a non-appropriate (and non-representative) sample
This action leads to results, applicable to a subset of the relevant population and not to the broader population itself. The chosen sample may seem to be representative of the relevant population, while it is not. The variability within the population may have not been taken into consideration, the size of the sample used may not be the appropriate or the method of selecting the sample may be uneven or limited [61].

Heuristics mapping: From a cognitive perspective, the *representativeness* and *availability* heuristic impact AI lifecycle at such a sampling step. For instance, it has been noted that during training, most AI face recognition algorithms, make use of pictures from people from developed countries, as people assume that these pictures are representative of the world population [41]. People consistently overestimate the proportion of the world that

is similar to themselves. This is why so much technology is optimized for able-bodied, white, cis-gender, males, because this population makes up the majority of developers [5]. Darker skin tones are underrepresented or even excluded, yielding less accurate results for dark skinned people. Moreover, already available or accessible datasets are usually preferred instead of the most suitable dataset. The *availability* heuristic rises due to the natural human tendency to place greater emphasis on the most readily available data, even if it does not fully represent the entire population [1, 27, 35]. The streetlight effect is a type of cognitive bias that arises when people only search for something where it is most convenient to look. As an example, researchers may assume that easily available, user-generated data from the Internet depicts reality, and is enough but this is not true.Large and easily available datasets do not always depict the world we live in.

Reflections: The resulting AI bias is the *representation* bias resulting in a model that fails to generalize well for certain subgroups of the relevant population. It is worth to mention that even given a perfect sampling, we can also face *historical* biases as datasets used may be available, and seem representative, but exhibit entrenched biases. Baked-in biases in the overall population (as results of people's bigotry) may regard disadvantaged groups including indigenous populations, women or disabled people and are perpetuated to the whole AI lifecycle, prompting discrimination[55]. As an example, CrimSAFE's algorithm was developed as a tenant screening algorithm to review and score potential renters in the United States. This algorithm disproportionately and illegally screened out Black and Latino applicants based on their criminal records. However, Black and Brown Americans are much more likely to be racially profiled, stopped and targeted by police for arrest [53]. This dataset reflected the already baked-in historical (also known as societal) biases, leading to a system that reinforced patterns of housing discrimination.

Harmful action 2: Misuse of proxy variables.

Usually a proxy, that is a concrete measurement, is chosen to approximate a construct i.e a concept, that is an unobservable or immeasurable variable. These proxy variables selected for a model can be poor reflections of target contrasts and/or be generated differently across groups. As an illustrative example, temperature may be used as a potential proxy for the "beauty" of the weather. This proxy is a poor reflection of the target construct. Ice cream sales are not dependent on the temperature but on sunshine [63]. In order for a variable to be a good proxy, it must have a close correlation with the variable of interest.

Heuristics mapping: A person who chooses which variables to include or avoid may decide in a way that aligns with their beliefs and impart their own cognitive biases into the model. The *representativeness* heuristic make people believe that a proxy is a good reflection of the true variable and is therefore used [9, 33]. A researcher may also use different measurement process across different groups due to *confirmation bias*. As an example, a fraud analyst might examine thoroughly some groups over others but these higher rates of testing will yield more positives and thus confirm the analyst's initial beliefs and prejudgements, skewing the observed base rates [64]. The *availability* heuristic can also lead to the use of a proxy that seems to relate to the inferred cause but such apparent correlation might be due to a third, hidden variable that is not available [41]. Illusory correlation (identified in Section 2 as an availability bias) makes an individual's perception of the relationship between two variables distorted, creating a false connection [62].

Reflections: *Measurement* bias occurs at this phase. Researchers may choose imperfect proxies for the true underlying values or include protected attributes and these actions result in inaccurate or/and discriminant classifiers [63]. Especially when it comes to protected attributes as proxy variables, unintended information about people can be revealed. Protected attributes partition a population into different groups that should, though, be treated equally. Developers avoid utilising protected attributes associated with social groups which have historically been discriminated against, as features in the training set. This would lead to a model that relies on sensitive information, yielding biased outcomes (*direct discrimination*). Removing the sensitive feature, though, does not prevent discrimination, as latent variables can be inferred implicitly e.g.gender can be inferred through browsing history, and race can be inferred through zip code [29, 55, 63]. *Indirect discrimination*, refers to cases

where groups or individuals are treated less favorably based on rules that disadvantage a protected group, while they seemed to be neutral. Redlining was a practice in the US, used to deny financial services in people based on their postal code. Neighborhoods that were supposed to be "risky" were outlined, but although postal code might seemed a neutral feature, it was highly correlated with ethnicity, leading to discrimination against African-Americans [64].

4.2 In-processing

This phase involves the model design, development, and evaluation. A model is trained to optimize a specified objective (e.g. minimize a loss function) and its quality is often measured on benchmarks i.e available datasets (different from the training data). What is needed is a model able to generalize on new, unseen data. The recent deployment of AI applications in social sensitive domains put forth a range of other important desiderata apart from accuracy, that models should be aiming for like fairness and privacy. Researchers use different benchmarks to estimate model's generalizability and select appropriate metrics and criteria that reflect the goals that had been initially set.

Harmful action 1: Make inappropriate model design choices.

Humans' model design choices regard aspects that range from the model's objective function to hyperparameter settings. What has to be understood is which model design choices disproportianately amplify error rates on features that may be underrepresented and possibly protected, raising fairness issues [29, 30].

Heuristics mapping: Decisions and who makes them at this phase can be driven by the *representation*, *availability* and *affect* heuristic. Information may be flattened as analysts seek patterns and try to come up with easier problems to solve even without being aware of this substitution (guided by representativeness and availability heuristics [11]. Certain objectives may rise as more important than others based on someone's existing beliefs and stereotypes that fill gaps in information. Consequently, choices give rise to harmful bias. Such choices are usually made in the absence of discriminatory intent. Simple models with fewer (and already available) parameters may be preferred, as they tend to be less expensive to build, more explainable, and transparent. By approximating a real-life problem, which is complicated, by a much simpler model, makes outputs easier to understand and this might be tempting, but introduces bias [51, 55, 62]. The choice of models' desiderata and hyperparameters is another matter of crucial importance at this phase. An analyst may believe that it is imperative to optimize for privacy guarantees in a health-care system based on their prior experience and inner beliefs, emphasizing the emotional importance of protecting personal information, but this can come at the cost of accuracy as discussed below [30].

Reflections: Several "flattening" techniques occur during inevitable mathematical abstractions leading to *algorithmic* bias. First of all, the choice of the model's objective function can lead to biased results for several sub-populations. Simple models fail to capture regularities in the data and tend to underfit the data. A model that does not generalize well for underrepresented groups causes harm and leads to discrimination for these underrepresented groups. In comparison, a model with high variance (and low statistical bias) may represent the data set accurately but could lead to overfitting to noisy or unrepresentative training data [30, 64].

Moreover, when it comes to the choice of a model's desiderata, there are several known trade-offs that have to be considered [19, 25, 30, 31, 68]. As an example, the loss function is used to evaluate how well the model fits the data. Loss on the training data has to be typically minimized, however, if the loss function is biased towards a specific group (e.g., white patients in a population), the relevant model will be better trained for this group. In this way, maximizing *accuracy* do not hold static other properties that are important like *fairness*[30, 67, 68]. As another example, differential *privacy*, which is the current state of the art in private machine learning, uses approaches like gradient clipping and noise injection. The cost of differential privacy, though, is a reduction in the model's accuracy and this can affect disproportionately underrepresented groups and model's *fairness*[2, 68].

Furthermore, error rates on features that are relatively underrepresented can get impacted disproportionately by even more subtle choices. Early stopping is a technique used to reduce overfitting without compromising model accuracy. However, underrepresented features are learnt later in the training process [30]. Model design choices should be made wisely to achieve an ideal balance between multiple model desiderata and avoid bias, as such choices definately express a preference for final model behavior.

Harmful action 2: Use of non-appropriate (and non-representative) benchmarks and inappropriate performance metrics.

During model evaluation, model's performance is measured on new and unseen data, usually called the test set. The main problem once again arises from a non-representative dataset that is now used to evaluate the model. The test data has to be different from the data used for training as the same dataset would lead to overoptimistic and overfitted models to a particular benchmark. The benchmark itself has to be representative of the population that the model will serve to avoid models that perform well only on a subset of the relevant population. Evaluation metrics are used to assess the performance and the effectiveness of a model.

Heuristics mapping: Usually, the model developer has to split the whole dataset into training and test data. Thus, the test dataset used may be once again non-representative of the population due to *representativeness* and *availability* heuristics as discussed in subsection 4.1. Empirical studies have also reported *confirmation bias* in software testers as a result of their tendency to design tests that confirm, rather than question the correct function of an algorithm. Testers may anchor on the belief that the algorithm is accurate and focus on scenarios or inputs that are consistent with this initial assumption, rather than exploring a broader range of possibilities that could challenge their initial belief [41, 54, 69].

Reflections: The resulting bias at this phase of the AI lifecycle is the *evaluation* bias. Choosing the wrong benchmark data can lead to overlook and perpetuate potential biases.As an example, a named entity recognition (NER) algorithm may be trained on a dataset with more female names as opposed to male names being tagged as non-person entities or not being tagged at all. In case a benchmark dataset primarily composed of male names is used, the NER algorithm will show good performance on it but experience disproportionately high error rates with female names that won't be even recognised as person enitities. However, in case a benchmark with more balanced gender representations is used, the misclassification of female names will be revealed.

Several metrics are used during a model's evaluation. Traditional model evaluation relies on common metrics like accuracy and precision to assess a model's performance. These metrics don't really explain if the model is fair to different groups of people, leading to unfair treatment that can be based on things like race, gender, or age. A subgroup validity approach is needed in this case, and protected attributes come now at handy to let analysts know the necessary group information (whether an individual belongs in a protected or unprotected group). Performance metrics are used across groups to avoid the use of one aggregated metric. Several fairness metrics are proposed in the literature but the plethora of fairness metric definitions illustrates that fairness cannot be reduced to a concise mathematical definition. The choice of the appropriate assessment criteria remains quiet cumbersome [40, 48, 49].

4.3 Post-processing

Post-processing steps regard the model's deployment and interpretation. There is also interaction of the AI model with the final users and this includes further learning and further development. The output of the model can be used as a new input to refine (e.g., through re-training) and (re)evaluate the algorithm creating what is known as "feedback loop". Without the feedback loops, AI systems would not be able to adapt to changing environments or improve their performance. However, AI algorithms can get influenced by their output, reinforcing and magnifying biases over time in several ways [47]. This self-perpetuating cycle of unfair decisions is resembled in literature to the butterfly effect [17]. Different types of biases are connected but understanding the dynamics

and the ways that various components are interconnected in such complex systems is quite cumbersome [17, 47]. Systems should undergo rigorous evaluation and monitored on an ongoing basis in the intended domain and the relevant risks have to be communicated [45, 55].

Harmful action 1: Use of a one-size-fits-all model.

This becomes harmful when a model is used for data in which there are underlying groups that should be considered differently. For instance, a model that is trained with a dataset composed of pictures of cats, dogs, and tigers, is used to predict the weight of the animal in an image. Using "dogs" or "felines" as labels can be misleading, since tigers and cats have different weights. A single model is unlikely to be best-suited for any group in the population, even if these groups are equally represented in the training data.

Heuristics mapping: Misinterpretation of statistical information is behind this action. Researchers collect statistical data in order to make generalizations from a sample to the population and not the other way around. When it comes to individuals, conclusions are drawn based on data collected at the group level. Stereotypes are usually claimed for a group, and then all members of this group are supposed to possess a set of essential characteristics. The *representativeness* heuristic makes people believe that a stereotype is true and all members of a group possess a number of essential characteristics [11, 32, 35]. *Conclusion* bias also affect this way of thinking, as people instead of doing the hard work of finding the best evidence, they just jump to a premature conclusion. Aggregated data often presents a simplified view, reducing complexity and cognitive load. When people rely on their feelings towards a group, they may overlook the diversity and individual differences within the group, resulting in incorrect assumptions about its individual members [20].

Reflections: The resulting bias at this phase of the AI lifecycle is the *aggregation* bias. Ecollogical fallacy is a cognitive bias in this category, that refers exactly to the inference that is made about an individual based on their membership in a group [4, 42].

Harmful action 2: Use of an AI model in ways not intended by developers.

During a system's performance validation it should be checked (among others) whether it is used in unintended ways or not[55]. For instance, a risk assessment model that is built to predict the likelihood of a criminal committing a crime in the future, should not be used to determine the length of a defendant's sentence [8]. Another common example in this category regards the distinction of AI systems between automated decision-making (ADM), where the system makes a decision automatically without any human involvement, and decision-support, where the system just supports a human decision-maker [45]. In case people put their faith in automated decision making whilst a model is not built for this, the consequences can be disastrous.

Heuristics mapping: It's human's cognition that should be understood and taken into consideration in this collaborative decision-making. *Anchoring* bias comes up when humans form a skewed perception due to an anchor, which is in this case the AI decision. As an example, when it comes to approaches that are more quantitative, subject matter experts may inadvertently activate a bias like the McNamara fallacy and take advantage of an AI system to take the pressure off of them in favor of the automation's objectivity [51]. *Confimation* bias could also be the reason for decision makers to favor suggestions from automated decision-making systems that align with their preexisting beliefs (and act as an anchor) and disregard decisions that contradict them, but could otherwise turn out to be more appropriate [26]. In case the suggestions align with their beliefs, the adjustment required is minimal, and the suggestions are easily accepted. Conversely, if the suggestions contradict their beliefs, the required adjustment is larger, leading to the dismissal or of these suggestions. Unfortunately, decision makers often use decision-support systems without fully understanding how they work. *Overconfidence* makes people overestimate their understanding of a domain but experts that make decisions are not necessarily familiar with machine learning, data science, or other fields associated with AI design and development. AI remains largely opaque, as advanced mathematics need to be understood by experts[24, 55]. The *affect* heuristic can also drive decisions at this phase. The more human-like a system appears, the more likely it is that users attribute more
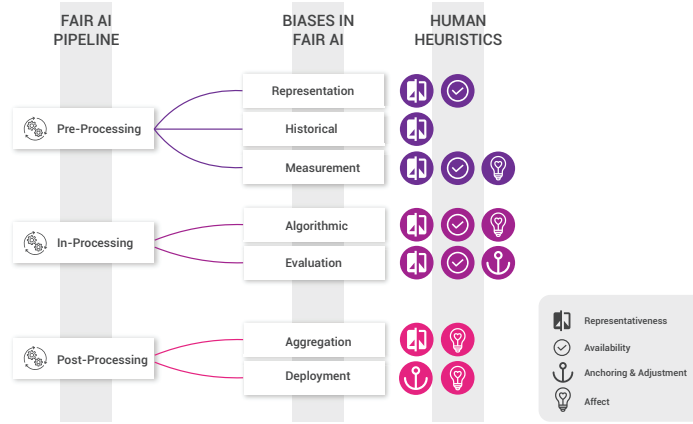
Fig. 1. Human Heuristics and Fair AI biases mapping

human traits and capabilities to it. Automated systems often create a sense of safety and comfort, reducing anxiety and leading to higher reliance. The emotional comfort provided by these systems can overshadow critical assessment. Anthropomorphising systems like conversational agents (CAs) is a common example of systems that lead to overreliance and unsafe use [65].

Reflections: The resulting bias at this phase is the *deployment* bias, that describes the discrepancy of how a model is used in comparison to what the model is designed to do. A system can be evaluated as if it were fully autonomous leading to automation bias that makes decision makers overestimate its credibility [45].

5 TOWARDS A HUMAN AND AI BIASES TAXONOMY - CONCLUSIONS

In our technology-laden lives, AI systems even carefully designed to be fair, remain biased with human-societal harms spotted over the AI lifecycle. Current computational and accuracy-based fair AI solutions fail to encompass sufficiently the full spectrum of risks posed by biases. Recognizing the necessity of a future where humans and AI work in synergy, this work offers a systematic mapping among the ways humans' heuristics lead to wrong or sub-optimal decisions which then are reflected in the AI detected computational biases.

Biases intertwine, causing cascading effects on the final AI system's performance. Biases in the collected datasets can lead to under-representation or exclusion of certain groups, resulting in discriminatory behaviors. An AI model,though, does not merely reflect existing biases in the data; humans' subjective choices regarding the act of selecting or computing features and model design choices during development can lead to measurement and algorithmic bias. Consequently, if a model is not assessed on a diverse dataset in real-world settings, some populations may be disproportionately impacted. Additionally, evaluation biases can emerge if inappropriate metrics are used. Finally, aggregation and deployment biases may arise if end-users are not adequately trained or supported regarding the right use of the model, further perpetuating a cycle of biased feedback loops.

These harmful actions stem from humans' heuristics that result in computational biases. The representativeness and availability heuristic seem to be responsible for the most types of computational biases like representation, measurement, algorithmic and evaluation bias. The representativeness heuristic is also responsible for actions that result from people's bigotry and result in historical bias. The role of affect, mood and emotion should also not be neglected. This automatic response from the intuitive system can generate judgements able to generate measurement, aggregation, deployment and even algorithmic bias. Anchoring and adjustment seems to be more manifest when a model is evaluated or deployed as an "anchor" becomes available.

Figure 1 illustrates a preliminary taxonomy of this heuristics-computational bias mapping, offering a new line of FairAI hypothesis setting and testing. With numerous cognitive biases that interfere with how people process data and think critically and several frameworks developed to identify computational biases in AI [38, 42, 61, 63], this work can't be exhaustive. Nonetheless, it constitutes a scientifically evidenced human and AI biases alignment summary, explanatory of the ways biases are perpetuated and amplified, and sheds light to prominent risks and vulnerabilities to consider when designing, developing, deploying, and evaluating FairAI solutions. Our analysis does not prove that other researchers' explanations or assumptions are wrong. Indeed, the biases in FairAI lifecycle may be a result of several factors, and even deeper study will need to explore the complicated and complex biases "iceberg". In summary, we claim that we offer a stepping stone to widen the study of human biases role in FairAI by revealing possible cognitive-computational biases intensities and interdependencies, largely uknown up to now. Our ongoing work on FairAI focuses on extending this mapping to a more holistic human-inclusive taxonomy which will be theoretically grounded and experimentatlly tested, under varying and bias-sensitive constraints posed in several critical FairAI domains.

REFERENCES

[1] Banerjee A and Nunan D. 2019. *Catalogue of Bias Collaboration.* Retrieved March 18, 2024 from www.catalogbias.org/availabilitybias

[2] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. *Differential privacy has disparate impact on model accuracy.* Curran Associates Inc., Red Hook, NY, USA.

[3] Max H. Bazerman and Don A. Moore. 1999. *Judgement in Managerial Decision Making.* John Wiley and Sons.

[4] Joel Best. 2021. *Is that true?: Critical thinking for sociologists.* Univ of California Press.

[5] Meredith Broussard. 2023. *More than a Glitch.* The MIT Press.

[6] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, Article 166 (2024), 38 pages. https://doi.org/10.1145/3616865

[7] Gabriela Cavagnolli, Ana Laura Pimentel, Priscila Aparecida Correa Freitas, Jorge Luiz Gross, and Joíza Lins Camargo. 2017. Effect of ethnicity on HbA1c levels in individuals without diabetes: Systematic review and meta-analysis. *PLOS ONE* 12, 2 (02 2017), 1–14. https://doi.org/10.1371/journal.pone.0171315

[8] Erin Collins. 2018. Punishing Risk. *Georgetown Law Journal* 57 (2018), 57–108. https://ssrn.com/abstract=3171053

[9] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The Measure and Mismeasure of Fairness. arXiv:1808.00023 [cs.CY]

[10] Daniel Fonseca Costa, Francisval de Melo Carvalho, Bruno César de Melo Moreira, and José Willer do Prado. 2017. Bibliometric analysis on the association between behavioral finance and decision making with cognitive biases such as overconfidence, anchoring effect and confirmation bias. *Scientometrics* 111, 3 (2017), 1775–1799. https://doi.org/10.1007/s11192-017-2371-5

[11] Kahneman Daniel and Frederick Shane. 2002. *Representativeness revisited: Attribute substitution in intuitive judgment.* Cambridge University Press, 49–81. https://doi.org/10.1017/CBO9780511808098.004

[12] Paul Slovic Daniel Kahneman and Amos Tversky (Eds.). 1982. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press.

[13] Virginia Dignum. 2023. Responsible Artificial Intelligence — From Principles to Practice: A Keynote at TheWebConf 2022. *SIGIR Forum* 56, 1, Article 3 (jan 2023), 6 pages. https://doi.org/10.1145/3582524.3582529

[14] Robert Dunne. 2016. Decision Making is Caused by Information Processing and Emotion: A Synthesis of Two Approaches to Explain the Phenomenon of Confirmation Bias. In *Irish Conference on Artificial Intelligence and Cognitive Science.* https://api.semanticscholar.org/CorpusID:15518030

[15] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: user-reported problems in intelligent everyday applications. In *Proceedings of the 24th International Conference on Intelligent User*

Interfaces (IUI '19). 96–106.

[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268. https://doi.org/10.1145/2783258.2783311

[17] Emilio Ferrara. 2024. The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications* 15 (March 2024), 100525. https://doi.org/10.1016/j.mlwa.2024.100525

[18] Emilio Ferrara. 2024. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* 6, 3 (2024). https://doi.org/10.3390/sci6010003

[19] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2023. SoK: Taming the Triangle – On the Interplays between Fairness, Interpretability and Privacy in Machine Learning. arXiv:2312.16191 [cs.LG]

[20] Slovic Paul Finucane Melissa L., Alhakami Ali and Johnson Stephen M. 2000. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making* 13, 1 (2000), 1–17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

[21] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (mar 2021), 136–143. https://doi.org/10.1145/3433949

[22] Paul Gallimore. 1996. Confirmation bias in the valuation process: a test for corroborating evidence. *Journal of Property Research* 13, 4 (1996), 261–273. https://doi.org/10.1080/095999196368781

[23] Dale Griffin Gilovich, Thomas and Daniel Kahneman (Eds.). 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

[24] Greg Githens. 2019. *How to Think Strategically: Sharpen Your Mind. Develop Your Competency. Contribute to Success*. Maven House.

[25] Alex Gittens, Bülent Yener, and Moti Yung. 2022. An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML. *IEEE Access* 10 (2022), 120850–120865. https://doi.org/10.1109/ACCESS.2022.3218715

[26] Kate Goddard, Abdul V. Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association : JAMIA* 19 1 (2012), 121–7. https://api.semanticscholar.org/CorpusID:2050860

[27] Rainer Greifeneder and Herbert Bless. 2007. Relying on Accessible Content Versus Accessibility Experiences: The Case of Processing Capacity. *Social Cognition* 25, 6 (2007), 853–881. https://doi.org/10.1521/soco.2007.25.6.853

[28] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://doi.org/10.1609/aaai.v32i1.11296

[29] Xiuting Gu, Zhu Tianqing, Jie Li, Tao Zhang, Wei Ren, and Kim-Kwang Raymond Choo. 2022. Privacy, accuracy, and model fairness trade-offs in federated learning. *Computers & Security* 122 (2022), 102907. https://doi.org/10.1016/j.cose.2022.102907

[30] Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem. *Patterns* 2, 4 (2021). https://doi.org/10.1016/j.patter.2021.100241

[31] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily L. Denton. 2020. Characterising Bias in Compressed Models. *ArXiv* abs/2010.03058 (2020). https://api.semanticscholar.org/CorpusID:222178157

[32] Nilsson Håkan. 2008. *In Search of Prototypes and Feminist Bank-Tellers: Exploring the Representativeness Heuristic*. Ph. D. Dissertation. Acta Universitatis Upsaliensis, Palo Alto, CA. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8476

[33] A. Cuevas J.R. Berrendero and J.L. Torrecilla. 2016. The mRMR variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation* 86, 5 (2016), 891–907. https://doi.org/10.1080/00949655.2015.1042378

[34] Tonni Das Jui and Pablo Rivas. 2024. Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics* (2024). https://doi.org/10.1007/s13042-023-02083-2

[35] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York: New York.

[36] Daniel Kahneman, Olivier Sibony, and R. Cass Sunstein. 20221. *Noise: A Flaw in Human Judgment*. Little, Brown Spark. https://doi.org/10.1007/s00283-022-10207-9

[37] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 1–6. https://doi.org/10.1109/IC4.2009.4909197

[38] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. The Pursuit of Fairness in Artificial Intelligence Models: A Survey. arXiv:2403.17333 [cs.AI] https://arxiv.org/abs/2403.17333

[39] Yeh W.-C. Lee C., Lee H.-Y. and Yu Z. 2022. The impacts of task complexity, overconfidence, confirmation bias, customer influence, and anchoring on variations in real estate valuations. *International Journal of Strategic Property Management.* 26, 2 (2022), 141–155. https://doi.org/DOI:https://doi.org/10.3846/ijspm.2022.16704

[40] Suvodeep Majumder, Joymallya Chakraborty, Gina R. Bai, Kathryn T. Stolee, and Tim Menzies. 2023. Fair Enough: Searching for Sufficient Measures of Fairness. *ACM Trans. Softw. Eng. Methodol.* 32, 6, Article 134 (sep 2023), 22 pages. https://doi.org/10.1145/3585006

[41] Naroa Martínez, Ujué Agudo, and Helena Matute. 2022. Human cognitive biases present in Artificial Intelligence. *Revista Internacional de los Estudios Vascos* 67, 2 (2022).

[42] Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54 (2019), 1 – 35. https://api.semanticscholar.org/CorpusID:201666566

[43] Patrick Mikalef, Kieran Conboy, Jenny Eriksson Lundström, and Aleš Popovič. 2022. Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems* 31, 3 (2022), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

[44] Rahul Mohanani, Iflaah Salman, Burak Turhan, Pilar Rodriguez, and Paul Ralph. 2020. Cognitive biases in software engineering: a systematic mapping study. *IEEE Transactions on Software Engineering* 46, 12 (1 Dec. 2020), 1318–1339. https://doi.org/10.1109/TSE.2018.2877759

[45] Information Commissioner's office. 2024. Annex A: Fairness in the AI lifecycle. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-i

[46] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, USA.

[47] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. 2023. A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23).* Association for Computing Machinery, New York, NY, USA, Article 7, 14 pages. https://doi.org/10.1145/3617694.3623227

[48] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. 2023. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing* 7, 1 (2023). https://doi.org/10.3390/bdcc7010015

[49] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (feb 2022), 44 pages. https://doi.org/10.1145/3494672

[50] Emily Pronin and Hazel Lori. 2023. Humans' Bias Blind Spot and Its Societal Significance. *Current Directions in Psychological Science* 32, 5 (2023), 402–409. https://doi.org/10.1177/09637214231178745

[51] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 83 (apr 2022), 22 pages. https://doi.org/10.1145/3512930

[52] J. Edward Russo and Paul Schoemaker. 1992. Managing Overconfidence. *Management Review* 33, 2 (1992), 7–17.

[53] Sara Safransky. 2020. Geographies of Algorithmic Violence: Redlining the Smart City. *International Journal of Urban and Regional Research* 44, 2 (2020), 200–218. https://doi.org/10.1111/1468-2427.12833

[54] Iflaah Salman, Burak Turhan, and Sira Vegas. 2019. A controlled experiment on time pressure and confirmation bias in functional software testing. *Empirical Software Engineering* 24, 4 (2019). https://doi.org/10.1007/s10664-018-9668-8

[55] Reva Schwartz, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270

[56] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19).* Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[57] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput. Surv.* 55, 13s, Article 293 (jul 2023), 39 pages. https://doi.org/10.1145/3588433

[58] Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. 2007. The affect heuristic. *European Journal of Operational Research* 177, 3 (2007), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

[59] Marco Smacchia, , and Stefano Za. 2023. Exploring Artificial Intelligence Bias, Fairness and Ethics in Organisation and Managerial Studies. In *ECIS 2023 Research Papers.* 362. https://aisel.aisnet.org/ecis2023_rp/362

[60] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive Biases in Fact-Checking and Their Countermeasures: A Review. *Information Processing & Management* 61, 3 (2024), 103672. https://doi.org/10.1016/j.ipm.2024.103672

[61] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (NY, USA) *(EAAMO '21).* Association for Computing Machinery, New York, NY, USA, Article 17, 9 pages. https://doi.org/10.1145/3465416.3483305

[62] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[63] Benjamin van Giffen, Dennis Herhausen, and Tobias Fahse. 2022. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research* 144 (2022), 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

[64] Hilde J. P. Weerts. 2021. An Introduction to Algorithmic Fairness. *ArXiv* abs/2105.05595 (2021). https://api.semanticscholar.org/CorpusID:234469664

[65] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[66] Alice Xiang. 2024. Mirror, Mirror, on the Wall, Who's the Fairest of Them All? *Daedalus* 153, 1 (03 2024), 250–267. https://doi.org/10.1162/daed_a_02058

[67] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A. Shenkman, Jiang Bian, and Fei Wang. 2022. Algorithmic fairness in computational medicine. *eBioMedicine* 84 (2022), 104250. https://doi.org/10.1016/j.ebiom.2022.104250

[68] Tao Zhang, Tianqing Zhu, Kun Gao, Wanlei Zhou, and Philip S. Yu. 2023. Balancing Learning Model Privacy, Fairness, and Accuracy With Early Stopping Criteria. *IEEE Transactions on Neural Networks and Learning Systems* 34, 9 (2023), 5557–5569. https://doi.org/10.1109/TNNLS.2021.3129592

[69] Gül Çalıklı and Ayşe Başar Bener. 2013. Influence of confirmation biases of developers on software quality: an empirical study. *Software Quality Journal* 21, 2 (2013). https://doi.org/10.1007/s11219-012-9180-0

This figure "acm-jdslogo.png" is available in "png" format from:

http://arxiv.org/ps/2407.21202v1