The Use of Non-epistemic Values to Account for Bias in Automated Decision Making

Jesse Hoey, Gabrielle Chan, Mathieu Doucet, Christopher Risi, and Freya Zhang*
University of Waterloo, Waterloo, ON, Canada N2L3G1
{jhoey,gabrielle.chan,mathieu.doucet,cjrisi,freya.zhang}@uwaterloo.ca

Abstract

We consider the algorithmic *shortlist* problem of how to rank a list of choices for a decision. As the choices on a ballot are as important as the votes themselves, the decisions of who to hire, who to insure, or who to admit, are directly dependent to who is considered, who is categorized, or who meets the threshold for admittance. We frame this problem as one requiring additional non-epistemic context that we use to *normalize* expected values, and propose a computational model for this context based on a social-psychological model of affect in social interactions.

1 Introduction

Definitions of algorithmic fairness include a subset that consider information beyond that included in a dataset. We propose here that such *"additional context"* can be found in human non-epistemic values, thought to be necessary for decision making [Mitchell et al., 2021, Friedler et al., 2019, Dotan, 2021]. An epistemic value is *"...one we have reason to believe will, if pursued, help toward the attainment of [such] knowledge"*, where *"[such]"* knowledge is the *"most secure knowledge available to us of the world we seek to understand"* [McMullin, 1982, p.18]. A non-epistemic value is everything but epistemic values. My neighbour, Chad, returns my garbage cans from the curb if I am absent. This is epistemic knowledge. The non-epistemic value associated with this is a societal expectation that neighbours *in general* are people who do helpful things. If measured, these non-epistemic values could potentially be removed from expected decisions, resolving bias.

There are three main problems in decision making: *agenda setting* (what options to consider), *framing* (how to understand the options), and *priming* (how to rank the options) [Scheufele and Tewksbury, 2007]. We consider the *agenda setting* problem, which is analogical to the decision of what to put on the menu in a restaurant. This problem is critical in many domains, such as democratic processes (who is on the ballot), hiring (who to interview), and marketing (what new populations to explore). In all these cases, what is on the "menu" has a significant impact on final decisions [Mercier, 2020, Chap. 9]. Further, shortlists are often made quickly and without rigorous justification, and so foster the manifestation of biases [Drage and Mackereth, 2022].

Consider the following example. A shortlist for a hiring decision is made using a mapping that ranks individuals with some irrelevant property ρ (call this set A) higher than individuals without property ρ (set B), but otherwise both A and B are expected to yield the same outcome, if hired. The bias about ρ leads to shortlist decisions that are Lipschitz unfair [Dwork et al., 2011], in that distributions over outcomes are larger for some group than would be expected. The degree of unfairness, however, could be estimated by asking a group of relevant people how they *feel* about each choice, assuming this question can be asked anonymously and the responses are honest. Suppose they feel comfortable with A being hired, but find that hiring B makes them uncomfortable, *independently of the outcome*

*JH,GC,CR and FZ are with the School of Computer Science, MD is with the Department of Philosophy. For more information see bayesact.ca

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(what job is being hired for). If the shortlist mapping functions are similarly biased, those with characteristic ρ should have ranking scores increased to offset the bias.

Beyond exposing this novel view of fairness, we also investigate the use of Affect Control Theory (ACT), a sociological model that estimates affective coherence based in language describing social events [Heise, 2007, 2010], as the non-epistemic values. ACT is a computational model that has been used to predict classes of human behaviour in a variety of settings [MacKinnon and Heise, 2010], including in moral decision making [MacKinnon, 2022].

2 Decision Problem

Decisions in which a person is evaluated are often framed as $\mathbb{E}(Q(d|v))$, where v are attributes of the person being evaluated, Q is the value of taking decision d when the person has attributes v, and d is the (typically binary) decision [Mitchell et al., 2021]. A sum over outcomes, o gives:

$$\mathbb{E}(Q(d|v)) = \sum\nolimits_o P(o|d,v)U(d,o), \tag{1}$$

where $U(d,o)$ is the utility of taking decision d and getting outcome o. In the terminology of Mitchell et al. [2021], person i has attributes v_i, and the decision about a target variable Y is made by estimating the conditional probability $P(Y=1|V=v_i)$. The outcome for person i is y_i, which is the same as o in Equation 1. The treatment in Mitchell et al. [2021] considers the utility function as simply an indicator of o, $U(d,o) = \mathbb{I}(Y=1)$ where $\mathbb{I}$ is the indicator function (so it essentially ignores U). Therefore, estimating $P(o|d,v)$ is all that is needed, using a scoring function $\psi(v_i)$.

Typical approaches then place emphasis on the features by splitting v in two parts, where x are unprotected features and a are protected features, and proceed by using $P(o|d,x)$ in place of $P(o|d,v)$.[2] However, many correlational "back-door" effects of a on o through the intermediary of x make this "unawareness" (of a) approach less than ideal in many cases. Instead, consider computing the expected value of even considering this person, which then involves a sum over decisions

$$\mathbb{E}(Q(v)) = \sum\nolimits_{d,o} P(d,o|v)U(d,o). \tag{2}$$

A shortlist is constructed by ranking individuals using Equation 2, and then applying a cut-off which is dependent of external factors such as how much time the committee has available. Suppose this filtering step aims to reject 90% of the individuals (high sensitivity at the cost of low specificity). The problem of selecting who to hire based on the shortlist is then relegated to a downstream process, one that involves *justifiable* rejection by humans of, say, 70% of the candidates interviewed (7% of the applicants), in order to hire 3% of the applicants. What is important here is that $P(d,o|v) \neq P(o|d,v)$. Since $P(d,o|v) = P(o|d,v)P(d|v)$, the usual method simply assumes that all decisions are equally likely when the outcomes are not considered, that is $P(d|v) = c$ with c constant. In fact, $P(d|v)$ is exactly where the bias lies.

In the following, suppose we can measure $P(d|v)$ using some population sampling tool (e.g. a survey or scrape of the web). For example, people could be asked if they would hire individual v *without knowing what was being hired for.* What this implies is a negative definition - we will define what *is not the case*. Fairness, when viewed as a lack of unfairness, is made up of differences between whatever you can measure that estimates $P(d|v)$. If decisions are different between v_1 and v_2 without considering outcomes, then this must be because of a bias about v_1 and v_2. This definition of fairness requires any differences in fair decisions between v_1 and v_2 to be based on attributes that are based solely on the problem being solved, e.g. the job being hired for. Any other features are relegated to $P(d|v)$, which we have assumed we can measure. Thus, knowing the history of individuals, decisions and outcomes (v, d, o, respectively), the inductively learned model of $P(o|d,v)$, call it $\phi(o,d,v)$, is assumed to contain two terms $\phi(o,d,v) = P^\dagger(o|d,v)P(d|v)$, where $P^\dagger$ is the expectation based on only epistemic factors. We can then use an estimate of $P(d|v)$, call it $\psi(d,v)$, to normalize $\mathbb{E}(Q(v))$:

$$\mathbb{E}(Q(v)) \approx \mathbb{E}_{norm}(Q(v)) = \sum\nolimits_{d,o} \frac{\phi(d,o|v)U(d,o)}{\psi(d,v)}. \tag{3}$$

Since we have defined unfairness to be exactly what can be measured with $\psi(d|v)$, the resulting ranking of individuals is generated only from $P^\dagger(o|d,v)$:

$$\mathbb{E}_{norm}(Q(v)) \approx \mathbb{E}_{FAIR}(Q(v)) \equiv \sum\nolimits_{d,o} P^\dagger(o|d,v)U(d,o), \tag{4}$$

[2] Protected features are ones on which a decision should not be based, such as race or gender.

Therefore, the type of bias measured by ψ is removed, such that a decision that neglects outcomes is independent of v. If all bias is centred around some particular variables, a, then they will be removed from $P^\dagger(o|d,v)$ through the normalization.

Therefore, our estimate of fairness requires an estimate of $P(d|v)$: decisions made by the population of decision makers about individual v, without considering outcomes. What we propose is a measure of how individuals *feel* about certain decisions regarding other individuals. For example, we can measure how hiring committees feel about hiring or not hiring person v *without considering what the job is, or the fit of* v. We have uncovered a bias if some type of person is not represented in the shortlist, and this would happen independently of the job due to historical experience or "back-door" effects. If we *define* bias precisely as decision differences when the outcomes are ignored, then all such bias is removed in $\psi(d,v)$, and we call the resulting decision ψ–fair (complement of ψ–bias). We consider an affective basis for ψ in the next section, although other interpretations are possible.

3 Affect Control Theory (ACT)

ACT is a social-psychological model of human social interactions based in sentiments about objects and events [Heise, 2007]. ACT maintains a static *denotative* model as an actor-behaviour-object state (e.g. *manager hires student*), and an associated connotative model: a dynamical system in Osgood's three-dimensional "EPA" space of affective meaning: evaluation (good vs. bad), potency (strong vs. weak) and activity (fast vs. slow). This dynamical system represents evaluative knowledge, whereas declarative and procedural knowledge are represented in the denotative model. The two models (denotative and connotative) are linked with a dictionary that maps from labels (e.g. *manager*) to EPA space. These sentiments are elicited using semantic differentials, in which individuals rate a word, say *manager*, on scales such as for evaluation with *good* at one end and *bad* at the other. Ratings are typically averages over about 1000 participants. The result for *manager* from the Indiana 2003 survey [Francis and Heise, 2006] is EPA:$1.0, 1.6, 1.3$.

Affective coherence in ACT is the difference between the sentiments elicited out of context, and the same sentiments elicited in a context given by an actor-behaviour-object triple representing a situation. This difference (squared) is called *deflection*, and measures how unlikely a given event is to occur. Thus, while *mother hugs child* is a low deflection (highly probable) event, *mother strikes child* is much higher in deflection (less likely). A key insight in this paper is that these deflections can be used as an independent measure of non-epistemic bias in decision making.[3]

We can construct a set of actor-behaviour-object events for a hiring decision with the deflections shown in Figure 1(b). Also shown are the deflections for the behaviour *fire-from-a-job* (EPA:$-1.1, 1.5, 0.4$). Thus, someone labelled *manager* would be more likely to *hire* (EPA:$1.7, 1.9, 1.1$) a *saleslady* (EPA:$0.6, -0.2, 0.6$) or a *student* (EPA:$1.5, 0.3, 0.8$) than a *criminal* (EPA:$-2.4, -0.8, 0.8$) or a *delinquent* (EPA: $-1.8, -0.8, 0.4$), indicating a bias in the population against criminals and delinquents. This bias will also be part of estimates, by the same population, of how successful each of these hires is. That is, the same population will rate delinquents as having a lower chance of success. More subtle differences, such as across gender, will yield smaller deflection differences. For example, the event *woman hire saleslady* has a deflection of 2.1, compared to 1.1 for *man hire saleslady.* By reversing the genders, we have uncovered a bias in the Indiana 2003 dataset [Francis and Heise, 2006].

To define $P(d|v)$, we use $\mathcal{D}$ for the deflection of a decision d. In this context, the actor is the decision-making body or committee, c, the behaviour is the decision, d (e.g. to hire or not hire), and the object is v, so the deflection is written $\mathcal{D}(c,d,v)$. With $\hat{\alpha}$ as an arbitrary scale factor, this is converted to a probability distribution (the ψ–fair ranking) following Hoey et al. [2021] as

$$P(d|v) \approx \psi(d,v) \propto e^{-\hat{\alpha}\times\mathcal{D}(c,d,v)}. \tag{5}$$

The assignment of labels to individuals and behaviours is a key component of this analysis. For example, the assignment of the label *manager* to someone may have to do with protected attributes. An applicant with some attribute facing a hiring committee biased against that attribute may be labelled as a *delinquent* while an applicant without that characteristic may be labelled as a *saleslady.* It is exactly this bias that we aim to remove by computing deflections. The assignment of labels to groups, however, is information which needs to be carefully elicited, see Hoey and Chan [2022].

[3]The numerical scores range between –4.3 and 4.3 for historical reasons. A deflection close to 1.0 is considered low, a probable event.

4 Exploratory Example - Hiring

Suppose we have two attributes: $r \in \{w, d\}$, which is protected, and $e \in \{m, b\}$, which we believe the decision making process should depend on (say this is a graduate or undergraduate degree). The population under study then have some (perhaps biased) $\phi(d, o, e, r)$ which they use to rank applicants for any setting of the two variables e, r. For each such setting, we can also construct an equivalent ACT event in actor-behaviour-object space as *"manager/self hires person of with attributes r, e."* Suppose that in some society, a bias against $r = d$ exists. Then simultaneously, one would expect that ϕ as measured in a population would decrease for $r = d$, while the deflection of the equivalent ACT event would increase. If the deflection for the event was still low, although ϕ was also low, then the normalized $P^{\dagger}(o|d, v)$ would stay the same: there is no bias against hiring this person, so the low ϕ estimate must be "real" and this person's ranking should remain low. The result is a ψ–fair decision-making algorithm in the sense defined in Section 3.

To quantify these notions, imagine that a measure of success, ϕ, *given* that a person is hired, is $\phi(o, d = \text{hire}, e, r)$ as shown in Figure 1(c), indicating a bias favouring people with $r = w$. Now consider deflection, and that the deflections are biased against persons with $r = d$ as follows. First, we have to assign identities to the different actors involved. Suppose we estimate that someone with $e = m, r = w$ will be labelled as a *saleslady* and someone with $e = b, r = w$ a *student*. However, due to a negative stereotype, persons with $e = m, r = d$ are labelled as *criminals* and those with $e = b, r = d$ as *delinquents*.[4] The deflections are those in Figure 1(b), and taking the exponent of the negative of this gives something proportional to the probability of success. The probabilities of *not hiring* also must be estimated using the event *manager fire [applicant]*.[5] This event, also shown in Figure 1(a), when normalized, gives us the final outcome probability. Repeating the process for a range of $\hat{\alpha}$ gives a set of unbiased probability matrices that we can compare to the original ϕ. To make the comparison, we look at (1) how inequitable it is across the protected attribute, shown as the KL-divergence of the distribution across the protected attribute, averaged over the unprotected one, shown in blue in Figure 1(a), and (2) how different it is from the original, shown as KL-divergences between the normalized and original distributions, averaged across all four conditions, shown in red in Figure 1(a). If we simply sum these,[6] we get a maximum at $\hat{\alpha} = 0.35$ (determined visually), which corresponds to the less biased probability matrix in Figure 1(d). So we can see the method is favouring equitable distributions, in particular over the $e = b$ class. Another way to interpret the blue curve in Figure 1 is as a measure of how far the distribution is from satisfying the Lipschitz condition above. After the inflexion point at $\hat{\alpha} = 0.35$ in Figure 1, the closest is achieved, and this is close to zero, after which the originally disfavoured group starts to gain disproportional advantage.

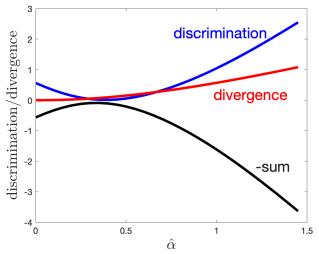
5 Limitations and Conclusion

We have described a method for normalizing automated decision making using non-epistemic values, and explored the use of this technique in a simple hiring scenario. This paper is largely conceptual, and has practical limitations which need more research. Key limitations are the external validity of the ACT surveys, multi-modal value distributions in affective space for polarizing identities, and the requirement to match the survey population with the decision making one. Further, the associations between applicants and identities in ACT was done manually in the context of the exploratory hiring example above. We show how these labels could be extracted from existing text corpora, present another example, consider the intersectionality problem [Kong, 2021], and discuss further limitations in [Hoey and Chan, 2022]. We stress the obvious importance of being careful with social engineering. We emphasize that our objective is for the algorithms proposed herein to be used strictly as informational devices for decision makers. A hiring decision could have an informed, unbiased reference point on which to base part of its decision, for example. A better understanding of bias in decision making can help in making automated decision making tools more easily explainable.

[4] Ideally, feelings about individual candidates would be measured directly, e.g. by averaging sentiments automatically measured in communications such as email and social media [Hoey and Chan, 2022].

[5] We use *fire* because *not hire* was not in the ACT survey used.

[6] How to balance these two is an empirical question based on the domain. Other notions of fairness beyond statistical parity, such as differential [Foulds et al., 2018] can also be used as measures of equity.



(a)

	applicant			
behavior	*saleslady*	*student*	*delinquent*	*criminal*
hire	1.1	1.1	3.2	4.1
fire	3.1	4.9	2.2	2.5

(b)

		unprotected	
		m	b
protected	w	0.9	0.7
	d	0.6	0.3

(c)

	unprotected	
	m	b
w	0.81	0.38
d	0.73	0.38

(d)

Figure 1: (a) Discrimination (blue) and model divergences (red) combine linearly to give (black) an best estimate of where both divergence and discrimination are minimized. The model predictions trade off equity (Lipschitz) with accuracy (how well they optimize the employer's loss function), but other weightings may also be possible. (b) The deflections for the event *manager* [*behaviour*] *applicant* using Indiana 2003 dataset [Francis and Heise, 2006]. The original (biased) estimate of ϕ is shown in (c), and final probability matrix shown in (d) for the maximum from (a) of $\hat{\alpha} = 0.35$.

Acknowledgments and Disclosure of Funding

We acknowledge funding from AGEWELL, Inc., Waterloo AI, and the Natural Sciences and Engineering Council of Canada (NSERC).

References

Ravit Dotan. Theory choice, non-epistemic values, and machine learning. *Synthese*, 198(11): 11081–11101, 2021.

Eleanor Drage and Kerry Mackereth. Does ai debias recruitment? race, gender, and ai's "eradication of difference". *Philosophy and Technology*, 35(89), 2022. doi: https://doi.org/10.1007/s13347-022-00543-1.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *arXiv*, 2011. doi: 10.48550/ARXIV.1104.3913.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. *arXiv*, 2018. doi: 10.48550/ARXIV.1807.08362.

Clare Francis and David R. Heise. Mean affective ratings of 1,500 concepts by indiana university undergraduates in 2002-3. Computer file, Distributed at Affect Control Theory Website, Program Interact bayesact.ca, 2006.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery.

David R. Heise. *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.

David R. Heise. *Surveying Cultures: Discovering Shared Conceptions and Sentiments*. Wiley, 2010.

Jesse Hoey and Gabrielle Chan. A novel approach to fairness in automated decision-making using affective normalization. *arXiv*, 2022. doi: 10.48550/ARXIV.2205.00819.

Jesse Hoey, Neil MacKinnon, and Tobias Schröder. Denotative and connotative control of uncertainty: A computational dual-process model. *Judgment and Decision Making*, 16(2):505–550, March 2021.

Youjin Kong. Intersectional fairness in AI? a critical analysis. In *Feminism, Social Justice, and AI Workshop as part of a special issue of Feminist Philosophy Quarterly.*, Waterloo, Canada (online), 2021.

Neil J. MacKinnon. Affect control theory applied to morality. *American Behavioral Scientist*, 0(0): 00027642211066042, 2022. doi: 10.1177/00027642211066042.

Neil J. MacKinnon and David R. Heise. *Self, identity and social institutions*. Palgrave and Macmillan, New York, NY, 2010.

Ernan McMullin. Values in science. *PSA: Proceedings of the biennial meeting of the philosophy of science association*, 1982(2):2–28, 1982.

Hugo Mercier. *Not Born Yesterday: The science of who we trust and what we believe*. Princeton University Press, 2020.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8: 141–163, 2021.

Dietram A Scheufele and David Tewksbury. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20, 2007.