




Statistical evidence and algorithmic decision-making

Sune Holm¹ 

Received: 8 February 2023 / Accepted: 24 June 2023 / Published online: 14 July 2023
© The Author(s) 2023

Abstract

The use of algorithms to support prediction-based decision-making is becoming commonplace in a range of domains including health, criminal justice, education, social services, lending, and hiring. An assumption governing such decisions is that there is a property Y such that individual a should be allocated resource R by decision-maker D if a is Y . When there is uncertainty about whether a is Y , algorithms may provide valuable decision support by accurately predicting whether a is Y on the basis of known features of a . Based on recent work on statistical evidence in epistemology this article presents an argument against relying exclusively on algorithmic predictions to allocate resources when they provide purely statistical evidence that a is Y . The article then responds to the objection that any evidence that will increase the proportion of correct decisions should be accepted as the basis for allocations regardless of its epistemic deficiency. Finally, some important practical aspects of the conclusion are considered.

Keywords AI · Statistical evidence · Fairness · Decision-making · Algorithm

1 Introduction

The use of algorithms to support decision-making is becoming commonplace in a range of domains including health, criminal justice, education, social services, lending, and hiring. For this reason, predictions made by algorithms are often viewed as “resource allocations awarded to individuals” (Hu & Chen, 2020, 535). For example, a bank may have a rule that a loan application should be approved if the applicant will not default. Not being able to observe whether an applicant is a to-be defaulter,

✉ Sune Holm
suneh@ifro.ku.dk

¹ Department of Food and Resource Economics, University of Copenhagen, Rolighedsvej 23, Frederiksberg, DK 1958, Denmark

the bank may deploy an algorithm to predict whether an applicant is a to-be defaulter based on a set of observable features such as employment status, credit history, and demographic information. The prediction is then used as evidence on which the decision is made. Call this scenario *Loan*.

Much of the debate about algorithmic decision-making focuses on what it means for an algorithm to be fair (Beigang, 2022; Binns, 2018a; Eva, 2022; Grant, 2023; Hedden, 2021, Holm 2022, Loi et al., 2021, Long, 2021). This article is about a related but distinct and comparatively neglected problem of algorithmic decision-making: Is it appropriate to allocate resources on the basis of purely statistical evidence provided by algorithmic predictions? Drawing on recent work on statistical evidence in epistemology, this article presents an argument against deploying algorithms to allocate resources when algorithms produce purely statistical evidence that a is Y . The article then responds to the objection that any evidence that will increase the proportion of correct decisions should be accepted as the basis for allocations regardless of its epistemic deficiency. Finally, some important practical aspects of the discussion are considered.

The argument of the article takes as its starting point the following case from legal philosophy:

Prisoner 100 prisoners are exercising in a prison yard. Suddenly 99 of them attack and kill the only guard on duty. One prisoner played no role whatsoever in the assault. These are the undisputed facts in the case, and there is no further information about what happened. If a prisoner is picked at random, his probability of guilt would be as high as 0.99 (Nesson, 1979).

Consider Hank, a prisoner who was in the prison yard. Given the evidence it is very likely that he is guilty of attacking the guard. How should we respond to this purely statistical evidence? Should we conclude that Hank was involved in murder and treat him accordingly? If we find Hank guilty based on the available evidence we will rely on *actuarial inference*: From information about the population-level frequency of some feature Y we infer the probability that “ a is Y ” is true, where a is an individual member of the population in question (Bolinger, 2021, 62). Is an actuarial inference about Hank an adequate basis for punishing Hank?

Most scholars find that we should not find Hank guilty based on purely statistical evidence. This is puzzling because we can make a variation of the case which generates the intuition that finding Hank guilty would be acceptable despite a decrease in the probability that he participated in the attack given the evidence. Consider the following variation on Prisoner:

Eyewitness You are presented with an eyewitness that testifies that Hank participated in the attack on the guard. You know that the eyewitness gets it right in 95% of cases.

It is generally assumed that we would accept the evidence provided in Eyewitness as sufficient to ground a guilty verdict. Moreover, we can conjure up cases where the purely statistical evidence provides an even higher probability that the defendant is liable than testimonial evidence, and yet the intuition persists: The testimonial evi-

dence in Eyewitness is sufficient for finding the defendant guilty, whereas the purely statistical evidence is not. What is going on here?¹

In response to pairs of cases like Prisoner and Eyewitness epistemologists have sought to explain our judgment that it is wrong to convict Hank in Prisoner, but not in Eyewitness. The explanations often point out that purely statistical evidence is epistemically deficient in a way that makes it inadequate to justify a verdict when the evidence, as in Prisoner, is purely statistical. Notably the criticism is not against the use of statistical generalizations and probabilistic inferences as such. Rather it is a criticism of the epistemic value of statistical evidence as *justification* for believing propositions of the form “*a* is *Y*.”²

Few people would accept punishing Hank based on the purely statistical evidence available in Prisoner. Should we also think that it would be inappropriate to allocate a resource to Hank based on purely statistical evidence? What if Hank applied for a loan and was rejected based on the statistical evidence that someone with his age, demographic features, and credit history is very likely to default? On the face of it, this seems like a relevantly similar case. I will argue that there is good reason to find statistical evidence equally problematic in the two cases. Allocative decisions are relevantly like other decisions about individuals, which we do not think should be based on purely statistical evidence. While I think clarifying how cases of algorithmic decision-making are relevantly similar to cases of court verdicts, this is not merely meant as a theoretical exercise. If my argument is valid, then it will have implications for the appropriateness of many of the envisaged uses of algorithmic decision-making.

The plan is this. In Sect. 2 I introduce algorithmic decision-making and the norm that governs it. Then, in Sect. 3, I present some illustrative cases of algorithmic decision-making, and provide a more formal outline of my argument. Taking Buchak’s (2014) analysis of blame as my starting point, Sect. 4 argues for the key claim that a decision-maker must be justified to *believe* that *a* is *Y* when deciding to allocate *R* to *a*. In Sect. 5, I present the case for thinking that the purely statistical evidence produced by algorithms does not justify belief that *a* is *Y*. Finally, in Sect. 6, I respond to an important objection to my epistemology-based rejection of algorithmic allocation. I conclude in Sect. 7.

2 The decision norm of algorithmic allocative decision-making

My focus is on algorithmic decision-making:

An entity that needs to make some decision—a decision-maker—defers to the output of an automated system, with little or no human input. These decisions affect individuals (decision-subjects) by conferring certain benefits or harms upon them. (Binns 2018b, 543).

¹ There is a huge literature on the puzzle of statistical evidence in law. See e.g., Redmayne (2008) and Gardiner (2018) for an overview.

² I return to some of the accounts on offer in Sect. 4.

I will assume that in all such cases of algorithmic decision-making, the decision must be based exclusively on the evidence afforded by the algorithm. While there are many contexts where the algorithmic evidence is just one piece of evidence available to the decision-maker, it is important to assess the value of the algorithmic input on its own. In part because the use of algorithms is widely expected to result in automation of more and more decisions. And in part because the insufficiency of exclusively algorithmic evidence for decision-making will highlight the need for drawing on other kinds of evidence.

Following Grant (2023, 3) the decision problem facing an algorithmic decision-maker can be described in the following way:

a decision-maker must decide whether to allocate some benefit or burden to particular individuals on the basis of whether they possess some feature that morally justifies allocating that benefit or burden to them. (Grant 2023, 3).³

Grant calls such decision problems “qualification problems” because there is a property Y , which determines whether it is “substantively fair” to allocate R to a . One of Grant’s examples of a qualification problem is deciding whether “a defendant in a criminal trial is innocent, and so ought to be acquitted” (Grant, 2023, 3). Adapting it to the case described earlier: If Hank is not guilty as charged, then he should be acquitted. Loan is also a case that can be characterized as a qualification problem. There is a property Y such that if loan applicant a is Y , then a should have her application approved. More generally, the sort of algorithmic decisions under consideration in this article are qualification problems, and thus they can be described as subject to the following assumption:⁴

The Merit Assumption There is a property Y such that if a is Y , then a should be allocated R .⁵

I will refer to Y as a “merit property.” Importantly, Y is a merit property relative to acceptance of a value for Y in the decision context. Thus, I do not want to claim that Y makes a merit R *simpliciter*. For example, some might argue that being male is a merit property, when loan applications are considered for approval. However, there might be very good reasons to reject this as the value of Y in the context of allocating loans. Still, in many contexts of applying algorithmic decision systems, there is consensus about the value of Y for a given R , and so it makes practical sense to consider the further question: Given that we have agreed on a value for Y in a decision context, does purely statistical evidence suffice to justify treating a as being Y ?

³ Let me note that while I rely on Grant’s characterization of the problem to which algorithmic decision-making is applied, I do not aim to discuss the merits of Grant’s argument for the Equality of Odds criterion of algorithmic fairness.

⁴ Grant lists several other examples of domains where qualification problems arise. Mitchell et al. (2021, 142) provides an even more comprehensive list of prediction-based decision-making in which “decisions are based on predictions of an outcome deemed relevant to the decision.” Thus, they remark that such techniques “are already common in lending, hiring, and online advertising, and increasingly figure into decisions regarding pretrial detention, immigration detention, child maltreatment screening, public health, and welfare eligibility.”

⁵ Thanks to an anonymous reviewer for pressing me to clarify this assumption.

Algorithms are typically introduced because, at the time of decision-making, decision-maker D cannot observe whether a is Y . Given the Merit Assumption, in such situations of uncertainty it seems appropriate to make a “best guess” about whether a is Y based on observable features of a . Thus, a responsible decision-maker will consider whether it is justified to decide *as if* a is Y given the available evidence. This gives us the following subjective norm for how to decide about whether to allocate R to a :

Subjective Decision Norm (SDN) D should allocate R to a if and only if D is justified to believe that a is Y .

The type of algorithms most often deployed as decision support are statistical models, trained on data about labelled historical cases (Corbett-Davies et al., 2017). When in use, the input to the algorithm is information about a set of observable features of an individual and the output is a score corresponding to the probability that *someone* with those features has the unobservable feature Y . For example, an algorithm may consider a loan applicant’s credit history, employment status, and marital status and on that basis make an accurate estimation of the probability that someone like the applicant will repay a loan (Verma & Rubin, 2018). For a *calibrated* algorithm an individual scoring 9 on a scale from 1 to 10 is an individual whose pattern of known features makes her belong to a group of individuals for which 9 out of 10 members are Y (Verma & Rubin, 2018). Thus, the risk score is based on assigning the individual to a reference group composed of individuals with a similar pattern of observable features. This is equivalent to assigning a score of 99 out of a 100 to Hank based on the observation that he was in the prison yard when the guard was attacked. In this way the evidence that a is Y provided by an algorithm is purely statistical.

3 Cases of algorithmic decision-making

In this article I focus on the use of algorithms to make binary classifications of individuals. To classify individuals, the algorithm applies a threshold such that individuals are classified as Y if their score is above the threshold. I will assume the threshold is set to maximize accuracy. Threshold-setting is an important topic in the literature on algorithmic fairness (see e.g., Long, 2021). However, for present purposes, the important point is that an individual is classified based on *the reference group* to which the individual is assigned by the algorithm. Importantly this means that such algorithms are using *actuarial inference* to classify individuals as Y or not- Y . Drawing on the available information they assign an individual to a reference group, and then they classify individuals based on their membership in that group.

Consider now the following case of algorithmic decision-making:

*Mistreatment*⁶ A county experiences an increase in referral calls for child mistreatment. To ensure that manual resources are directed towards calls which report a genu-

⁶ Something like the scenario described here has already been implemented in several countries in Europe and in the US. See e.g., Chouldechova et al. (2018) for a case study. The scenario presented is inspired by this case, and cases from European countries as well. In the case discussed by Chouldechova et al. (2018), the system is intended as decision-support to personnel receiving and screening referral calls. I present

ine problem, the county deploys an algorithm which can accurately predict whether a referral call is about an actual case of mistreatment. If the algorithm predicts a call to be genuine, then a case worker is assigned to make an in-person investigation. Calls about children that are classified as innocuous are not further investigated but screened out.

To illustrate, imagine that a call comes in about Hannah. The algorithm accesses information available in public records about Hannah and on the basis of the observable features it assigns Hannah's case a risk score of 3 out of 10. This means that 3 out of 10 children, who are like Hannah in their observable features, suffer mistreatment. Hence, Hannah's case is screened out and no in-person investigation is initiated.

Like Prisoner and Loan, Mistreatment is a case in which a decision is being made about an individual based on the probability that a proposition of the form " a is Y " is true. However, in Mistreatment, as in Prisoner, the evidence for the truth of the proposition is purely statistical. Thus, if we think that the evidence in Prisoner is insufficient for punishing Hank, then it seems we should also think that the evidence for finding Hannah "guilty" of not being maltreated is insufficient. Similar cases can be described for other domains of decision-making. Here's one from a health context:

Retinopathy In country X, the national health agency has decided that, due to a lack of ophthalmologists, diabetes patients should be pre-screened for diabetic retinopathy (DR), an eye disease that if untreated will cause blindness. Healthcare professionals at local health clinics are asked to deploy a system which calculates a risk score for patients based on images of their retinas. It then applies a threshold to classify the patient as suffering or not suffering DR. The system has been tested to achieve the same level of accuracy as human experts. Those classified as suffering DR are recommended for specialist examination. Those classified as negative for DR are asked to come back for rescreening in 12 months.⁷

We may imagine that Harry, who is a diabetic, scores 2 on a scale from 1 to 10 and is therefore classified as negative for DR and hence not allocated specialist examination. The basis for this allocation will be purely statistical evidence and actuarial inference. The frequency of DR in patients whose retinal images display the sort of pattern displayed by images of Harry's retinas is 2 in 10.

Decision-making under uncertainty may also take place in everyday settings such as in a situation where one finds that one's phone has been stolen:

*Stolen Phone*⁸ You leave the seminar room to get a drink, and you come back to find that your mobile phone has been stolen. There were only two people in the room, Jake and Barbara. You have no evidence about who stole the phone, and you don't know either party very well, but you know (let's say) that men are 10 times more likely to steal phones than women.

In this case the purely statistical evidence makes it highly likely that it is Jake and not Barbara who stole your phone. Still, Buchak remarks,

it as a case of automated screening of such calls to keep focus on the value of the evidence provided by the algorithm.

⁷ This case is based on the use of IDx-DR as described in FDA (2018). It might seem obvious that this is a case in which we should simply maximize accuracy. I will return to discuss this case in Sect. 6.

⁸ See Buchak (2014, 292).

(...) this isn't enough to make you rationally believe that Jake stole the phone. If you accused Jake, he could, it seems to me, rightly point out that you don't have evidence that he in particular stole the phone. He could protest that you only know something about men in general or on average. (Buchak 2014, 292).

What Buchak points out here, on behalf of Jake, is that an actuarial inference from purely statistical evidence does not justify a belief that an individual person has a certain property *Y*. This is in effect a claim that the evidence available does not suffice for deciding to treat Jake as if he is *Y* because SDN is not satisfied: You are not justified in believing that Jake stole your phone, so you should not act towards him as if he had.

Relatedly, Buchak points out that it would be inappropriate for you to *blame* Jake for stealing your phone. This is because according to the "blame norm" you should blame someone for being *Y* if and only if you are justified in believing that they are *Y*.⁹ However, you are not justified in believing that Jake stole your phone in *Stolen Phone* (2014, 299). Thus, Buchak can be understood as suggesting that our practice of blaming is governed by SDN.

My argument in this article concerns decisions to which SDN applies. Decisions to punish and to blame individuals seem to be governed by SDN. And so does the allocative decisions in *Loan*, *Mistreatment*, and *Retinopathy*. Epistemologists have used cases like *Prisoner* and *Stolen Phone* to argue that statistical evidence is insufficient for punishment and blame because it does not justify belief that *a* is *Y*. I want to argue that similar considerations support the view that statistical evidence is insufficient for allocation decisions more generally. Like decisions to punish and to blame, allocative decisions should be based on evidence meeting a certain evidential standard which is not met by purely statistical evidence.¹⁰ More formally, my argument can be stated in the following way:

1. If *D* should allocate *R* to *a* if *a* is *Y*, then *Y* is a merit property with respect to allocation of *R* by *D* to *a*.
2. If *Y* is a merit property with respect to allocation of *R* by *D* to *a*, then *D* must be justified to believe that *a* is *Y* given *D*'s evidence, when deciding to allocate *R* to *a*.
3. If *D* only has purely statistical evidence that *a* is *Y*, then *D* is not justified to believe that *a* is *Y* given *D*'s evidence.
4. If algorithmic decision-makers (ADs) only have purely statistical evidence that *a* is *Y*, then ADs cannot meet SDN.
5. Hence, ADs should not make allocative decisions on the basis of mere statistical evidence.

⁹ In Buchak's words: "Blame someone if and only if you believe (or know) that she transgressed" (2014, 299).

¹⁰ Enoch et al. (2012, 210) argue that it is "utterly implausible" to sacrifice accuracy in order to live up to some epistemological standard. To them, the reasons for not relying on purely statistical evidence in court are instrumental. My argument is an attempt at describing why it is that punishment, blame, and allocative decision procedures in general are not merely a matter of accuracy, i.e., minimizing erroneous decisions. I return to Enoch et al.'s argument in Sect. 6.

Call this the Merit Argument. If the Merit Argument is sound, then it presents a fundamental challenge to the introduction of algorithmic allocative decision-making. It also provides an explanation of why many people feel uneasy about deferring important allocation decisions to algorithmic predictions. The unease can, at least in part, be explained as arising from the recognition that the evidence on which the decision is based is epistemically deficient. In the next section I consider the case for (1) and (2).

4 Merit and belief

Let me first present my reasons for accepting (1). To begin with, consider a situation of deciding under certainty. There is a population of loan applicants and Y is a property we can observe. For the sake of argument, let's say that Y is *has a full-time job*. In this situation, it seems clear that applicants with a full-time job should be approved for a loan because they have the merit property. Moreover, an applicant with a full-time job, who is not approved for a loan, has a complaint. In other words, when making allocations under certainty about whether a is Y , it seems clear that a merits R if a is Y .

Now, imagine that the decision must be made under uncertainty. We do not know whether applicants have a full-time job. However, we do know something about their education and demographic features, so we can produce an algorithm that can accurately predict whether applicants have a full-time job - say with 95% accuracy. Now, the reason we want to predict whether applicants have a full-time job seems to be that this is what makes them merit a loan. We don't predict whether they play tennis. Why? Because it does not matter for whether they should have the loan. In conditions of uncertainty, we *still* think that it is being Y that merits an allocation of R . However, when we don't know whether a is Y , the best we can do is to make a "best guess" about it. An accurate predictive algorithm can be a very good tool for making such guesses in the sense that it can provide highly probative evidence. However, it is not the case that in situations of uncertainty the evidence that a is Y comes to constitute a 's merit or claim to R .¹¹

To see this, consider a situation in which we have a file with information about whether loan applicants have a full-time job. Consulting the file, we annotate which applicants should be approved for a loan. Unfortunately, the computer with the information is damaged before the loans are paid out, so we must assess applicants all over again. This time we do not have direct access to information about their job situation. However, we can use an algorithm to make accurate predictions. In this situation, the evidence on which we base our decision has changed to become purely statistical, but whether an applicant merits a loan has not changed from being a matter of whether the applicant has a full-time job to being the likelihood that the applicant has a full-time job. Now, according to my argument, this is the case for other allocative

¹¹ My argument here is based on Broome's (1984) argument against the view that we should find merit in the equality of expected utilities just because we think that there is merit in equalizing the distribution of a given total of utility.

decisions governed by SDN as well. It is *the children* suffering mistreatment and *the patients* with DR, who merit the resource of specialist examination and in-person assessment. In short, Y is a merit property in that if *a* is Y, then *a* merits R regardless of whether we know that *a* is Y.

Still, why think that a responsible decision-maker should *believe* that *a* is Y when allocating R to *a*? Is it not enough for the decision-maker to have a sufficiently high rational credence that *a* is Y? To answer this question, I will expand on Buchak's analysis of the role of belief in contexts of blaming (Buchak, 2014, 288, see also Littlejohn, 2020).

To begin with Buchak observes that our practice of holding each other responsible essentially involves the expression of reactive attitudes such as resentment, blame, and gratitude. And while these reactive attitudes come in degrees, "the degree of blame I assign to a particular agent is based on the severity of the act, not on my credence that she in fact did it" (Buchak, 2014, 299). In other words, when I appropriately blame Jane to some degree for stealing my bike, the degree of blame is not a function of how likely it is that Jane stole my bike given my evidence. Rather I blame her in proportion to the act of stealing my bike because I believe that she stole my bike.

Commenting on the norm for expressing reactive attitudes such as blame Buchak notes that:

(...) reactive attitudes associated with blame, like resentment and indignation, are *partially constituted by representing the world as being such that their targets are culpable for the act.* (Buchak 2014, 308, italics added).¹²

Thus, according to Buchak's analysis, blaming someone for an action involves two independent judgments. One judgment is "settling on what the world is like" (2014, 309). This is an epistemic component. Another judgment is "determining the amount of blame that is appropriate when the world is like that" (2014, 309). This is a value component. Thus, when blaming under uncertainty about what the world is like, i.e., about whether blame candidate *a* has merit property Y, we require that one is justified in believing that *a* is Y, i.e., in believing that *a* is blameworthy. It is being Y that makes an individual blameworthy, not the likelihood of being Y. Hence, our evidence for finding *a* to be Y must meet a certain epistemic standard. It must justify belief that *a* is Y.

Summing up, my support for premise (2) is this: Both blaming and allocating are subject to SDN. As shown by Buchak, it is very plausible to think that blaming is a value judgment based on an independent yes-no judgment about whether *a* is blameworthy (is Y). I have argued for a similar analysis of allocative decisions. Allocating R to *a* is a value judgment about whether *a* should have R based on an independent yes-no judgment about whether *a* is meritorious (is Y). In the next section I consider

¹² In fact, it might be that in addition to reactive attitudes other emotions are also subject to a norm according to which they are appropriate if and only if warranted by a justified belief. Thus, Buchak surmises that "fear is appropriate when and only when you justifiably believe something dangerous" (2014, 308).

whether purely statistical evidence should be considered sufficient justification for belief that a is Y .

5 Belief and evidence

I now turn to the third premise of my argument namely that purely statistical evidence does not justify belief. If true, this means that, in so far as algorithmic evidence is purely statistical, it does not suffice for making allocative decisions governed by SDN. Still, one might argue that a belief that a is Y is justified when the evidence makes it sufficiently likely that the belief is true. Thus, one might propose:

Rational Threshold View There is a threshold t such that D is justified to believe that a is Y if and only if the probability that a is Y given the evidence is above t .¹³

If we accept the Rational Threshold View, we can argue that in Prisoner it is justified to believe that *Hank is guilty* if and only if the evidence makes one's rational credence that *Hank is guilty* sufficiently high. Whether the evidence is statistical or non-statistical is irrelevant for belief according to this view. Hence, on the Rational Threshold View, it can be perfectly rational to believe that *Hank is guilty* on the grounds of purely statistical evidence. A similar line of reasoning will apply to the cases of using algorithmic evidence as the basis for allocations. If it is sufficiently likely that Hannah is not being mistreated or that Harry does not have DR, then a decision-maker is justified in believing that they do not merit the resource being allocated, regardless of the purely statistical nature of the evidence.

For the purpose of this article suffice it to say that the Rational Threshold View faces significant objections and that it is widely accepted that "bare statistical evidence cannot produce belief" (Buchak, 2014, 292). First, it entails that our responses to the Prisoner and Eyewitness cases are inconsistent. If we think it justified to believe that Hank is guilty based on the eyewitness testimony, then any kind of evidence that makes it equally or more likely that Hank is guilty should also justify belief that he attacked the guard.

Second, the Rational Threshold View also seems to give us the wrong answers to lottery cases. Imagine that you have a ticket for a lottery with one million tickets. While it is extremely improbable that your ticket is a winner, and perfectly rational for you to have an extremely high credence that it is a loser, it will not be rational for you to believe that *the ticket is a loser*. After all, you know that you never know whether it is a winner.¹⁴ Thus, we should recognize that whether the evidence justifies a belief is not exclusively a matter of the *probability* of the truth of the belief conditional on the evidence. It matters what kind of evidence we have.¹⁵ In Smith's words,

¹³ See Buchak (2014, 289) for a similar formulation.

¹⁴ See e.g., Hawthorne (2004): "(...) the epistemic subject under consideration has good reason for being confident that the lottery proposition is true—the lottery proposition is highly likely relative to the person's evidence. And yet I take it as a datum that there is a strong inclination to claim that the relevant lottery propositions are not known. Nor is this merely a datum about the inclinations of philosophers. After all, the motto of the New York State lottery is 'Hey, you never know'."

¹⁵ More can be said on behalf of the Threshold View than I do here. It is not my aim to present a thorough discussion of the view, but simply to point to the central arguments that make it widely rejected by episte-

we need “more than probability, but less than certainty” (Smith, 2016, 4) when it comes to the justification of belief.

Epistemologists have devised a variety of accounts of what it is that is missing in cases where we find it unjustified to believe a proposition P on the basis of purely statistical evidence. Smith’s own suggestion is that statistical evidence is insufficient for justifying belief because it lacks “normic support.” On this account,

(...) a body of evidence E normically supports a proposition P just in case the circumstance in which E is true and P is false requires more explanation than the circumstance in which E and P are both true. (Smith 2016, 40).

For example, if my evidence that the proposition *the screen is green* is true is based on my perceptual experience of the screen, then it would, in normal circumstances, be surprising if the proposition was false. Surprising in the sense that I would look for an explanation for why my perceptual evidence misled me. However, if my evidence that the proposition *the screen is green* is my knowledge that the color of the screen is determined by a random lottery with odds 999:1000 that the screen is green and odds 1:1000 that it is blue, then I would not be surprised and look for an explanation in case it turned out that the proposition *the screen is green* was false.

Like Smith’s normic support account other accounts also find that statistical evidence is lacking in an epistemically important respect. According to the *sensitivity* account, what is missing in cases of purely statistical evidence is sensitivity of the evidence to the falsity of the proposition. Consider Eyewitness. The eyewitness’ testimony that she saw Hank attacking the guard would be false if Hank did not attack the guard. However, the statistical evidence in Prisoner would still be true. Thus, it is not sensitive to the truth of the proposition for which it serves as evidence (see Enoch et al., 2012).

According to the *safety* account, the reason why purely statistical evidence does not justify belief is that it is “unsafe.” To say that evidence E is safe with respect to P is to say that there are no close possible worlds in which P is true and E is false (Pardo, 2018). However, in cases of purely statistical evidence, a possible world where the evidence is true, and P is false does not have to be remote at all. For example, it does not seem as if a possible world where Hank is innocent and 99% of prisoners in the yard are guilty must be very different from the actual one. In other words, we could easily be wrong when relying on purely statistical evidence.

Finally, and most recently, Jackson (2020) has suggested a helpful distinction between two kinds of evidence which explains why statistical evidence does not justify rational belief and tells us something about what sort of evidence can justify rational belief (Jackson, 2020, 5083–5084):

B-evidence Evidence for P that does *not* make salient the possibility of not-P.

C-evidence Evidence for P that makes salient the possibility of not-P.

Based on Jackson’s distinction, we can explain why purely statistical evidence does not justify rational belief that P: It amounts to mere C-evidence. On its own

mologists. For further discussion see e.g., Fassio and Gao (2020) who argue that a modified version of the Rational Threshold View can avoid the problem of statistical evidence.

statistical evidence for P is such that for an agent responding appropriately to the evidence the possibility that not-P will be salient. To illustrate, consider Prisoner. If the only evidence a jury member has for finding that *Hank participated in the assault* is that Hank was in the prison yard, when the assault took place and that 99% of the prisoners in the prison yard participated in the assault, then the possibility that Hank did not participate in the assault will be salient to the juror (Jackson, 2020, 5085). After all, the juror cannot properly appreciate the evidence and *not* pay attention to the fact that one of the prisoners is innocent, and that Hank might be that prisoner.

In general, Jackson suggests, when we believe a proposition to be true this is due to the possession of evidence which does not make it salient for us to consider the possibility that the proposition is false. Still, as Jackson also notes, this does not entail that belief that P requires that not-P is not salient. In fact, as is often the case, a juror might believe that the defendant is guilty while being well-aware that the belief might be wrong. Still, the *evidence* that justifies the juror's belief in the guilt of the defendant must include sufficient B-evidence to justify the belief. The purely statistical evidence about Hank does not amount to such a justification.¹⁶

It is not my aim to assess the merits of different accounts of why purely statistical evidence is lacking when it comes to justifying belief. However, when taken together they present a strong case for thinking that such evidence cannot justify belief. Still, one might ask: Why care about such epistemological niceties when making allocative decisions? If we have a resource and must make decisions about who should get it under uncertainty, shouldn't we simply aim for maximal accuracy – for the highest proportion of correct decisions?¹⁷ If we care about giving resources to those who merit them, then we should aim to get it right as much as we can and not be obsessed with the epistemic value of the evidence we rely on. I consider this view in the next section.

Summing up, a premise of the argument in this article is that purely statistical evidence does not justify belief that *a* is Y. In this section I have presented the case for this claim by outlining recent work in epistemology on statistical evidence. My point is that, in their own way, each of these accounts point to an explanation of why it is that statistical evidence is not sufficient for justified belief that *a* is Y.

This concludes my defense of premises (1)-(4) in the Merit Argument. The argument is basically that because allocations should track people's merit, they must be based on evidence that can justify belief. In the next section I present and respond to an important objection to this claim.

¹⁶ Jackson is aware that her account entails that there are cases in which the testimony of an eyewitness can be presented such that it counts as C-evidence. In response Jackson points out that her account concerns non-ideal, rational agents and how they should respond to the evidence. For such agents "how evidence is presented can make a difference to the appropriate doxastic response" (2020, 5089).

¹⁷ This sort of question is posed forcefully by Enoch et al. (2012).

6 What about accuracy?

I have suggested that epistemology should matter for allocation decisions. Such decisions reflect a judgment about whether an individual merits a resource and therefore purely statistical evidence is inadequate as the sole basis for making such decisions. However, an obvious objection to this view is that an assessment of the appropriateness of an allocation decision procedure should be its accuracy, not its epistemology. This has been argued forcefully in the context of the use of statistical evidence in legal verdicts. In this section I present and respond to this argument.

Enoch et al. (2012) presents an important criticism of the view that the law should care about epistemology. Their claim is that when it comes to making verdicts, we should not dismiss types of evidence which would increase the accuracy of the verdicts. To drive home their point, they present the following thought experiment:

Suppose you have to choose the (criminal) legal system under which your children will live, and you can choose only between systems A and B. System A is epistemologically better: perhaps its courts only convict when they know (or think that they know) the accused is guilty, or perhaps they only convict based on sensitive evidence, or perhaps they convict only based on evidence that normically supports the conclusion that the accused is guilty. System B is not as good epistemically as System A. But System B is more accurate, so that the chances of System B convicting an innocent are lower than the chances of System A doing so. Which system do you choose for your children: the Epistemologically-Fine-But-Not-That-Reliable System A, or the More-Reliable-But-Not-That-Epistemically-Respectable System B? (Enoch et al. 2012, 209).¹⁸

Enoch et al. (2012) finds that we would prefer System B over System A. We may consider similar thought experiments with respect to Loan, Retinopathy, and Mistreatment. Would we want a procedure for allocating loans, specialist examination, and in-person investigation to be like System A or System B? The conclusion of the Merit Argument seems to be that we should accept System A even if there is an algorithmic alternative that will increase the proportion of correct decisions.

Interestingly, Enoch et al. acknowledge that while

epistemological considerations, (...), never *by themselves* defeat considerations of accuracy. (...), it is possible that epistemological considerations defeat considerations of accuracy *indirectly*, via some other considerations to which they are relevant. (Enoch et al. 2012, 211).

.Thus, a legal finding that Hank is guilty may stand in a “close normative connection” to a moral attitude such as blame, and therefore such a finding may be inappropriate

¹⁸ Pundik (2011) presents a similar line of argument concluding that “If there is any justification for restricting the use of statistical evidence in court, it does not lie in epistemology; it has to lie elsewhere” (2011, 143).

if the evidence for it does not ground the accompanying moral attitude. However, Enoch et al. do not pursue this indirect approach further.

The Merit Argument amounts to an indirect argument for the significance of epistemology for allocation decisions. It argues that the legitimacy of allocative decisions is not simply a matter of them being produced by the most accurate procedure available. Their legitimacy requires that the procedure is subject to a procedural constraint, namely that they rely on a certain kind of evidence.¹⁹ My argument can thus be said to make an indirect case for the relevance of epistemology in allocative decisions-making. It is relevant because in order for the decision procedure to be legitimate it must rely on evidence which is not purely statistical. Allocative decisions are, on this analysis, expressions of an evaluation of whether an individual merits a resource. And to be legitimate, i.e., to be such that individuals should accept and conform to them, they must rely on the right kind of evidence.

In contrast, on the accuracy approach appropriate allocation decisions are not based on “settling what the world is like”, but on “betting what the world is like.” To maximize accuracy the decision-maker considers whether a is sufficiently likely to be Y for it to be a rational bet to allocate R to a . The threshold for when a candidate for R is sufficiently likely is determined by where to set it to maximize (expected) accuracy. On this approach, the legitimacy of an allocative decision procedure is exclusively a matter of its accuracy. There are no constraints on the procedure by which decisions are made, except for constraints that will improve accuracy. Thus, on this approach, any kind of probative evidence should be used.

The accuracy approach underlies common arguments for applying algorithmic decisions-making. Algorithms should be deployed because they can provide us with more (or at least no less) accurate decisions than alternative human-involving decision-making. And they can do so faster and for more people thereby increasing the efficiency of the decision-making. However, the common rejoinder to this argument is that something is missing when people are subjected to algorithmic decisions. This is often expressed in terms of a concern that people are not treated as individuals. I have proposed that there is an epistemological dimension to this kind of concern. Deciding about whether an individual should have a resource expresses a judgment about their merit, and hence they must be based on evidence that can justify belief about their merit.

7 Concluding remarks

In this article I have tried to connect recent work on statistical evidence in epistemology and recent work on the use of algorithmic decision-making. I have argued that we should reject algorithmic allocative decisions when they will be based on purely statistical evidence. I then considered the objection that rejecting algorithmic decision-making will likely come at a cost in accuracy.

¹⁹ Another constraint on allocation procedures is fairness. E.g., the maximally accurate decision procedure may not be acceptable because it has a disparate impact on salient groups. For discussion of the costs of fairness constraints see e.g., Corbett-Davies et al. (2017).

If my argument is accepted this has significant practical implications for the use of algorithmic predictions as the basis for allocative decisions. For example, it would seem to exclude the use of automated algorithmic decision-making to make pre-screening decisions in cases such as Retinopathy. I find this conclusion hard to accept. However, I also think that this is not the conclusion that should be drawn from my argument. Let me try to explain why.

What I think my discussion brings out is that when we introduce algorithmic decision-making, we should do so with open eyes. If we apply it in the way envisaged e.g., in Retinopathy, then we must be explicit about the moral justification for doing so. We must show why we think that the gain in efficiency is morally better, all things considered, than using a procedure that is epistemically adequate. In this way I suggest that the epistemic constraint on allocative decision-making supported by the Merit Argument is not absolute. It should be considered as a *pro tanto* reason not to rely on purely statistical evidence. However, there are other values at stake when we consider the legitimacy of a decision-making procedure. Accuracy is one of them. Fairness is another. When implementing algorithmic allocation procedures, we must be transparent about how we justify the degree to which the procedure accommodates these different values.

Funding Open access funding provided by Royal Library, Copenhagen University Library.

Declarations

Conflict of interest The author has no conflicts of interest to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Beigang, F. (2022). On the advantages of distinguishing between Predictive and Allocative Fairness in Algorithmic decision-making. *Minds & Machines*, 32, 655–682. <https://doi.org/10.1007/s11023-022-09615-9>.
- Binns, R. (2018a). Fairness in machine learning: Lessons from political philosophy. Conference on Fairness, Accountability and Transparency. PMLR. Retrieved June 3, 2023, from <https://oa.mg/work/2963808661>.
- Binns, R. (2018b). Algorithmic accountability and public reason. *Philosophy and Technology*, 31, 543–556. <https://doi.org/10.1007/s13347-017-0263-5>.
- Bolinger, R. (2021). Explaining the Justificatory asymmetry between statistical and individualized evidence. In J. Robson, & Z. Hoskins (Eds.), *The Social Epistemology of legal trials* (pp. 60–76). Routledge.
- Broome, J. (1984). Uncertainty and fairness. *The Economic Journal*, 94, 624–632.
- Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies*, 169, 285–311. <https://doi.org/10.1007/s11098-013-0182-y>.

- Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency in Proceedings of Machine Learning Research*, 81, 134–148. Retrieved June 20, 2023, from <https://proceedings.mlr.press/v81/chouldechova18a.html>.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. Retrieved June 20, 2023, from <https://doi.org/10.1145/3097983.3098095>.
- Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs*, 40, 197–224.
- Eva, B. (2022). Algorithmic Fairness and Base Rate Tracking. *Philosophy and Public Affairs*, 50, 239–266.
- Fassio, D., & Gao, J. (2020). Belief, credence and statistical evidence. *Theoria*, 86, 500–527. <https://doi.org/10.1111/theo.12261>.
- FDA (2018). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. Retrieved June 20, 2023, from <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye>.
- Gardiner, G. (2018). Legal burdens of proof and statistical evidence. In D. Coady, & J. Chase (Eds.), *The Routledge handbook of applied epistemology* (pp. 179–195). Routledge.
- Grant, D. G. (2023). Equalized odds is a requirement of algorithmic fairness. *Synthese*, 201, 101. <https://doi.org/10.1007/s11229-023-04054-0>.
- Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.
- Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49, 209–231.
- Hu, L., & Chen, Y. (2020). Fair Classification and Social Welfare. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain*. ACM, New York, NY, USA. <https://doi.org/10.1145/3351095.3372857>.
- Jackson, E. (2020). Belief, credence, and evidence. *Synthese*, 197, 5073–5092.
- Littlejohn, C. (2020). Truth, knowledge, and the standard of proof in criminal law. *Synthese*, 197, 5253–5286. <https://doi.org/10.1007/s11229-017-1608-4>.
- Loi, M., Herlitz, A., & Heidari, H. (2021). Fair Equality of Chances for Prediction-based Decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 756. <https://doi.org/10.1145/3461702.3462613>.
- Long, R. (2021). Fairness in Machine Learning: Against false positive rate Equality as a measure of Fairness. *Journal of Moral Philosophy*, 19, 49–78. <https://doi.org/10.1163/17455243-20213439>.
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–163. <https://doi.org/10.1146/annurev-statistics-042720-125902>.
- Nesson, C. (1979). Reasonable doubt and permissive inferences: The Value of Complexity. *Harvard Law Review*, 92, 1187–1225.
- Pardo, M. S. (2018). Safety vs. sensitivity: Possible worlds and the law of evidence. *Legal Theory*, 24, 50–75.
- Pundik, A. (2011). The epistemology of statistical evidence. *The International Journal of Evidence & Proof*, 15, 117–143. <https://doi.org/10.1350/ijep.2011.15.2.373>.
- Redmayne, M. (2008). Exploring the Proof Paradoxes. *Legal Theory*, 14, 281–309.
- Smith, M. (2016). *Between probability and certainty: What justifies belief*. Oxford University Press.
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)* Association for Computing Machinery, New York, NY, USA, 1–7. <https://doi.org/10.1145/3194770.3194776>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.