Cambridge Quarterly of Healthcare Ethics (2024), 33: 3, 370–379
doi:10.1017/S0963180123000646

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Healthy Mistrust: Medical Black Box Algorithms, Epistemic Authority, and Preemptionism

Andreas Wolkenstein

Institute of Ethics, History and Theory of Medicine, Ludwig-Maximilians-Universität (LMU) München, Lessingstr. 3, D-80336 Munich, Germany
Email: andreas.wolkenstein@med.uni-muenchen.de

Abstract

In the ethics of algorithms, a specifically *epistemological* analysis is rarely undertaken in order to gain a critique (or a defense) of the handling of or trust in medical black box algorithms (BBAs). This article aims to begin to fill this research gap. Specifically, the thesis is examined according to which such algorithms are regarded as epistemic authorities (EAs) and that the results of a medical algorithm must completely replace other convictions that patients have (*preemptionism*). If this were true, it would be a reason to distrust medical BBAs. First, the author describes what EAs are and why BBAs can be considered EAs. Then, preemptionism will be outlined and criticized as an answer to the question of how to deal with an EA. The discussion leads to some requirements for dealing with a BBA as an EA.

Keywords: Black box algorithms; trust; ethics of algorithms; epistemology; preemptionism

Introduction

In the case of so-called black box algorithms (BBAs), many people—scientists, politicians, and laypeople alike—adopt a skeptical attitude of mistrust.[1] One of the main reasons is the epistemic opacity of these algorithms. In the following, the term "epistemic opacity" will refer to a "cognitive mismatch between the complex mathematical operations performed by ML algorithms and the type of reasoning used by human beings"[2]. It is about the fundamental incomprehensibility of those decisionmaking processes that lead the algorithm to link a certain input with the corresponding output. BBAs are therefore "computers that write their own programs"[3]. Furthermore, I will use the term "transparency" in order to refer to all and any kinds or measures aimed at reducing the level of opacity that is deemed problematic.

The fact that we cannot comprehend the process of belief generation seems to be a problem for many people.[4] It is primarily *legal and ethical challenges*, for example, regarding the attribution of responsibility or the proof of discrimination, that are at the forefront of this debate.[5] Jens Christian Bjerring and Jacob Busch argue that black box medicine, that is, medical practice that uses BBA for diagnostic and therapeutic decisions, poses a threat to the ideal of patient-centered decision-making.[6] Specific *epistemological* challenges are also sometimes cited when critically assessing BBAs. According to Juan Manuel Duran and Karin Rolanda Jongsma, opacity in the case of medical BBAs raises serious moral concerns. They argue "that certain actions are morally unjustified given the lack of the epistemic warrants required for the action to take place. A physician is not morally justified in giving a certain treatment to a patient unless the physician has reliable knowledge that the treatment is likely to benefit the patient"[7].

An important difference between ethical and epistemological problems is that certain measures are sufficient to solve the former, but not necessarily the latter sort of problems. For example, even if the potential for discrimination in BBAs could be minimized through detailed ex ante or ex post audits, there

would still be the problem that a doctor's actions in relying on an unintelligible instrument are not justified. Duran and Jongsma state in this sense: "While these moral concerns [e.g. discrimination; author's note] are genuine, they neglect the epistemological bases that are their conditions of possibility. We propose, instead, that the epistemology of algorithms is prior to, and at the basis of studies on the ethics of algorithms."[8]

The fact that epistemological problems remain, even if responsibility and non-discrimination are clarified, is easy to understand, especially in the medical (ethical) field. For example, the principle of autonomy, which is considered one of the four central ethical principles for identifying and solving ethical problems in medicine, could be used.[9] From this, it could be argued, it follows that patients must always be fully informed in order to give their informed consent to treatment. If the doctor cannot explain how her diagnosis, which was made by a BBA, was arrived at, then she would be violating the autonomy requirement.[10] Admittedly, a discussion would have to develop in the solution of this problem, pointing for instance to the great healing potential of successful algorithms, which is stronger than the autonomy requirement. Nevertheless, the violation of the autonomy principle remains, which makes the use of BBAs appear problematic in the long term, especially when autonomy is given a superior position.

Following on from this, it could also be argued that a *sui generis* problem exists when beliefs are exchanged (and acted upon) whose genesis is fundamentally incomprehensible. Perhaps it is to be accepted as a problem in its own right that people experience discomfort when they are treated on the basis of beliefs that cannot be understood, whether in the medical field or elsewhere. Possibly there is some kind of epistemic minimum requirement for the justifiable (non-)acceptance of beliefs, for example, basic insightfulness or understandability. This could be justified, similar to the medical field, with the principle of epistemic autonomy: Only if all (necessary) information is available—and this includes basic comprehensibility—does the acceptance of a belief contradict this ideal.

What exactly is the reason that we refuse to trust or accept machine-generated beliefs? And is this mistrust justified? Based on the premise that the epistemology of algorithms takes precedence over an ethical consideration, an explicitly epistemological perspective is adopted in the following. This means that the question is pursued as to whether the cause and a justification for the distrust can be found in dealing with BBAs as epistemic *authorities* (EAs). The fact that BBAs can be considered EAs is a recurring theme though not explicitly mentioned; the fact alone that we defer to BBAs because of their capacity to provide diagnoses proves this point. Often, it is concluded that trust is necessary for the attribution of EA to BBA, but that this is not justified for either conceptual or ethical reasons.[11] Others raise a skeptical voice and argue that there is more to trustworthy algorithms than merely ensuring their validity and effectiveness.[12] Rarely, however, is a specifically *epistemological* analysis of the attribution of EA undertaken in order to gain a critique (or a defense) of trust in BBAs. This article aims to begin to fill this research gap.

Specifically, the thesis is examined according to which the conviction of an EA must completely replace other convictions that laypeople have (*preemptionism*). If this were true, it would be a reason to distrust EAs, because such a requirement contradicts ideals such as epistemic autonomy. Accordingly, the further procedure is as follows: First, I will describe what EAs are in the first place and why BBAs can be considered EAs. Then, preemptionism will be outlined and criticized as an answer to the question of how to deal with an EA. This is followed by the argument to be examined here, according to which preemptionism (or the discomfort with it) justifies the non-acceptance of EAs. This discussion in turn leads to some requirements for dealing with BBAs as EAs.

What is Epistemic Authority? And are Black Box Algorithms Epistemic Authorities?

We encounter EAs in everyday life in the most diverse areas: In questions for which we have no answers, we turn to people (but also institutions, media, and collectives)[13] who we assume are in an epistemically better position than we are. That is, we assume they have more and better reasons to make a certain decision, and they can thereby answer our questions. And because they are in this position, it is very likely

that we will obtain a true answer to our questions and consequently have another true belief in our belief system if we adopt the opinion of authority.

Christoph Jäger defines EA as follows:[14]

> EA_{RF}: A is a (recognized) epistemic authority for S in D at t and relative to G, iff S truly believes A to be able, and prepared, to help S achieving S's epistemic goals in D and with respect to G, where this ability is due to A being in a substantially advanced epistemic position in D, relative to S, at or during t, and with respect to the goal of acquiring G.

To speak of an EA thus involves at least two persons or bodies, authority A and person S, a certain domain of knowledge D, a temporal reference t (because EAs are not necessarily permanent for S), and an epistemic goal or good G (such as knowledge, true conviction, or understanding that A wants to attain). Furthermore, EA consists not least in the fact that S recognizes A as EA, but that this recognition is actually true. Moreover, EA has a pragmatic aspect: To be an EA requires certain abilities and skills to help A achieve the corresponding epistemic goal G. And finally, this ability is grounded in the superior epistemic position in which the EA finds itself.

What is disputed is how EA relates to epistemic expertise: Are experts necessarily EAs? Jäger denies this: Neither is expertise necessary for being an EA, nor is expertise sufficient: EA does not depend on an absolute state of knowledge, but is always given relative to the state of knowledge of a subject.

EAs are thus those persons, institutions, or collectives to whom we correctly attribute an epistemically superior position in a domain of knowledge and who can successfully help us achieve our epistemic goal (knowledge, true beliefs, or understanding). Experts can also be EAs, but *as experts* they show different skills than EAs who guide the novice to understand a fact or a domain of knowledge. Methodologically, she can draw on a variety of measures, from informing to questioning to maieutics.

The conclusion that we regard BBAs as EAs follows from the abovementioned conceptual considerations. After all, they are developed and used precisely to provide us with knowledge or at least true beliefs, because and insofar as they have more true beliefs than a doctor or even the patient herself. Their increasing use in medicine is evidence that they have authority in a particular domain of knowledge (radiology, for example) and this is mainly epistemic—they create true beliefs in us and do so because they have excellent means (if they are good) to achieve the epistemic goal of obtaining true beliefs about the patient's condition. Algorithms, if they are good, are in a better epistemic position than the doctor or the patient. So, there is nothing to be said against describing dealing with BBAs as dealing with EAs. Is this a reason to distrust BBAs? In order to answer this question, it is necessary to analyze how the beliefs produced by an EA are dealt with.

Dealing with EA: Preemptionism

Two major approaches to this have emerged in the literature. The first, *preemptionism*, requires novices to adopt the fact that an EA has a certain belief (that p) as the sole reason for having that belief (that p).[15] This fact thus *replaces* other reasons that a person has. In contrast, the *total evidence view* approach requires that the fact that an EA has a belief (that p) is included in the totality of reasons that a person has regarding the particular issue. When EAs are invoked, there is an aggregation of reasons and beliefs in S, a balancing.

Preemptionism has a certain intuitive plausibility, especially in situations where (a) a novice has no prior beliefs or reasons, or where (b) the novice has conflicting beliefs or reasons. In case (a), preemptionism is almost trivial because, strictly speaking, there is nothing that needs to be replaced; the expert belief simply takes the place of the reasons guiding the novice's belief. Elizabeth Fricker speaks of "weak deferential acceptance," when an EA offers beliefs and the novice has no opinion or is not inclined to generate a belief either, should she reflect on the relevant question.[16] And in case (b), by adopting the expert's opinion, we seem to increase the chances of finding the truth, at least if one assumes an objectivist understanding of EA (according to which an EA is distinguished by actually having a

higher number of true beliefs in the area in question). Here, Fricker speaks of "strong deferential acceptance," that is, when novices themselves have certain beliefs, but the authority is in an epistemically better position to elicit the decisive reason for or against a belief.

Moreover, preemptionism gains plausibility through its connection with the concept of epistemic trust. Benjamin McMyler argues that trust is the goal of EA: EA requires us to set aside our other reasons and accept the EA's belief; this is what epistemic trust would consist of.[17] And according to Arnon Keren, trust implies lowering one's epistemic safeguards; that is, the trusting party refrains from taking actions to counter possible adverse effects should the trust in the speaker turn out to be false.[18] Jäger summarizes:

> "By contrast, keeping one's own reasons in play would amount to taking epistemic precautions against acquiring a false belief from the speaker. Epistemic trusters abandon such additional epistemic precautions, hence they preempt."[19]

Distrust of BBA as Fear of Preemption

With these considerations in mind, we can now address the question of what (the lack of) trust in BBAs is all about. First of all, we could characterize the mistrust that many people have of BBAs as epistemic mistrust. One way to situate this distrust more precisely would be to interpret it as *fear of preemption*: To trust epistemically implies to accept the fact that an EA has an opinion as the exclusive reason for my opinion, where this reason supersedes all my other reasons regarding a specific question (e.g., the cause for the patient's symptoms). This, however, speaks against the ideal of (epistemic) autonomy, which we consider enormously important, especially in medicine. Last but not least, informed consent is a concept that has developed in medical practice and is often seen as a central component of medical ethics. The principle of autonomy is considered by quite a few to be the most important principle of medical ethics, and its instrument for implementation—informed consent—therefore acquires essential importance.

This interpretation of mistrust as at least partly caused by the fear of having one's own beliefs and reasons for beliefs replaced by a belief generated by an algorithm as an EA is reinforced by the considerations of the trust argument (see above). Insofar as a BBA represents an EA, we are called upon to drop our epistemic safeguards—and this is precisely the goal of EA according to the trust argument. This, however, seems irresponsible to us, especially in the case of a diagnosis. After all, it is sometimes a matter of life or death, or at least a not inconsiderable intervention in our lives. A lot depends on the correct diagnosis. The diagnosis itself can change our lives and sometimes causes a great deal of suffering. Being confronted with the news that one has lung cancer will rarely have no effect on the patient's psychological state. The question of treatment also depends on the correct diagnosis. A benign carcinoma may be removed surgically—possibly inconveniently, but with only a short duration of damage—whereas a malignant carcinoma requires radiation or chemotherapy. Here, the consequences are much more protracted and problematic. However, should preemptionism present itself not just as a description but as a normative demand on us in dealing with EAs, we refrain from trusting BBAs in cases where they demand from us to replace our beliefs with algorithm-generated diagnoses even in live-or-death decisions.

One might argue that preemptionism is in and of itself an untenable and unrealistic position. Indeed, this is what the discussion below will reveal. However, as an interpretation of what people assume, preemptionism serves as a helpful tool. Even though I do not make an empirical claim, that is, I am not saying that preemptionism is empirically proven to be the motivation for people's mistrust, I think there are a few hints that could be used to consider (fear of) preemptionism regarding BBAs as the driving force behind the widespread and observable mistrust. *Firstly*, technology often has a massive impact on people's lives in the sense that the introduction of technology makes its use almost imperative. Speaking epistemologically, this could mean that people might feel obligated to use medical algorithms as the final expert word in a given diagnosis. This might be even more the case when all that people hear is that medical algorithms are better than humans. They might feel pushed toward accepting the algorithm's diagnosis as the belief that ought to replace their prior convictions. *Secondly*, when BBAs become ubiquitous and patients meet physicians who solely or at least heavily rely on algorithms to make

diagnoses, algorithms whose inner working mechanisms they do not really understand, patients might come to believe that since there is no one who could challenge the algorithm's result, there is simply no other option than to let the algorithm's diagnosis replace all other beliefs. After all, what options would they have other than not believing or believing? Depending on the severity of the illness, not believing a diagnosis could have disastrous consequences. And while acting on a false-positive result might also be detrimental to patients' well-being, choosing the treatment in this case could be justified as the risk-averse option. *Thirdly*, time constraints and other resource shortages (financial or otherwise) in the healthcare sector might lead to a situation where any kind of challenging the algorithm's results is disadvised. In other words, there could simply be no room for challenging and contestation so that eventually and effectively people would have strong incentives or even would be more or less coerced to replace their beliefs by the algorithm's results.

Although these factors are to be found in the way new technology assumes its place in society and the specific conditions of the healthcare sector, there is a *fourth* element that might contribute to fear of preemption as the driving factor for mistrust against BBAs. It has to do with the specific capacities that pervade medical algorithms and particularly medical artificial intelligence (AI). Intelligent algorithms do not merely follow rules—mechanical, statistical, or otherwise—that have been "programmed" into them by humans. In these cases, it might be a lot easier to reject potential mistrust by referring to their reliability. Medical AI acts, as does all AI, much more autonomously and comes with a huge amount of credibility in terms of reliability. Thus, we face not simply a more sophisticated tool that physicians use as much as they use blood tests or functional magnetic resonance imaging (fMRI). As will be briefly discussed below, physicians using a BBA might be considered not as physicians using a tool but as physicians consulting another expert. And in cases like this, where the consulted expert is a powerfully knowing one, it seems much more understandable that we feel the need to accept what this expert has to say. Essentially, the reason is that the algorithm does not simply provide a test result as an output, but offers a diagnosis, that is, a belief that one might feel compelled to accept as the final word. Whereas test results are an *indicator* of a diagnosis, waiting for confirmation by a physician, the BBA offers the full picture which is, many people feel, not in need of further confirmation or interpretation. This might be at the core of why preemption is an available and recommended reaction to the use of algorithms in medicine. Expressing mistrust, in turn, is a reaction to this perceived pressure to let one's beliefs be replaced by the all-powerful machine.

From all that has been said before, there is now an argument that could be used to show that the mistrust is unjustified.

(1) Distrust in algorithms is (caused and) justified by the assumption of preemptionism, that is, the assumption that we should let all our reasons be replaced by the fact that a BBA p says.
(2) But preemptionism is false, that is, we do not have to let the fact that a BBA says p supersede all our other reasons.
(3) Therefore, distrust of BBA is not justified.

That (1) is true makes intuitive sense because preemptionism is against the ideal of epistemic autonomy and (therefore) strikes us as irresponsible.

Note that it is, at this point, not argued that preemptionism is false *because* it is incompatible with the ideal of autonomy. What is said instead is that autonomy requires that we refrain from replacing our beliefs with an algorithm-as-EA's belief so that whenever we face an algorithm-as-EA, we should be cautious. But this has, at first sight, nothing to do with preemptionism itself, but with the concrete situation where, in a life-or-death decision, it would be somewhat irresponsible to let a black-box-generated belief replace all our other beliefs. Put differently, it might be the case that, epistemically speaking, preemptionism is true *and* it is incompatible with the ideal of autonomy, at least in certain circumstances. But it could be still true, as will an examination of premise (2) show, that preemptionism is wrong or implausible also or partly on grounds of considerations that have to do with epistemic autonomy. So, what is the state of preemptionism?

Is Preemptionism Tenable?

Critics have argued that preemptionism does not provide an answer to the question of when it is rational to lower one's epistemic safeguards. The trust approach merely states that when trust is made, it consists in lowering the safeguards. Moreover, it seems to be the case that trust is attributed gradually. A listener can trust an EA even if she only partially lowers her safeguards; a complete "dropping," that is, a total preemption of her own reasons, is not necessary.

It is also questionable whether the fact that the EA says *p* simply supersedes our other reasons or whether it is not rather the case that the balance of pro and con reasons regarding *p* changes in p's favor. If preemptionism were correct, there would be unacceptable consequences in some cases: For instance, if novices empirically falsify an EA belief, one would have to argue that the EA loses its authority over the novice because the empirical falsification is outside the EA's sphere of action.[20] One better explains such cases with a rejection of preemptionism.

Preemptionism (at least in its strong variant) is made even more implausible by the following considerations: First, in cases where there is an initial agreement between the beliefs of EA and novice, the epistemic situation of the latter seems to improve by adding the EA belief and even worsen by preemptionism. This is true even, as Dormandy has shown, if one assumes that the novice's reasons are already included in the EA's reasoning (against the *double counting* argument).[21] And Jäger writes:

> "Suppose that we base our belief that *p* on a non-conclusive reason *r* and learn that A has the same belief at least partly for the same reason. In that case, anti-preemptionists argue, what S can rationally add to *r* is the belief that A believes, at least partly on the basis of r, that p. This complex reason is not identical with r, and it is hard to see why it could not rationally be aggregated with *r* itself and supply additional support of p."[22]

It seems that preemptionism can capture some intuitions about our epistemic practices in dealing with EA. But it does not seem to be relevant to all such practices. Rather, the critiques suggest that we should retain a non-negligible degree of suspicion so as not to drop our complete epistemic safeguards, for instance. This also seems more compatible with the ideal of epistemic autonomy. This demands that we should only hold beliefs on the condition that we ourselves have thought about the truth of these beliefs. And should we not be able to do this, then at least, one would conclude from what has been said before, not all epistemic safeguards should be dropped. We do this by retaining a measure of suspicion and not letting the beliefs of the EA (or the fact that an EA has a certain belief) completely replace our own beliefs and reasons. This also becomes clear with regard to the dangers attached to an excess of trust (*credibility excess*), as Jäger points out.[23] Moreover, this seems to open the way to a procedural understanding of EA, a doxastic practice as an expression of EA.[24]

The following points of criticism are particularly important for the context of BBA:

a) Epistemically responsible is an aggregation of evidence, not a preemption; but here a distinction must be made: Is there further evidence (e.g., is there a medical opinion found without the BBA, or is there otherwise no prior belief)? And if there is such prior evidence: Is it consistent or inconsistent with the result of the BBA?
b) A certain degree of epistemic precaution is necessary to avoid, for example, *credibility excess,* and this is quite compatible with the attribution of trust, which is attributed gradually.
c) Moreover, the demand to drop one's epistemic safeguards seems to be too imprecise. For strong trust-based preemptionism, there seems to be only an either-or: Either one has epistemic safeguards or, if and insofar as one trusts, one drops them. But obviously the boundary, the *threshold* at which one is inclined to drop one's safeguards, changes. Although it seems to be no problem to trust quickly in this sense for simpler questions, we demand more from our counterpart when the *stakes* are high; when the risks involved in trusting are high; and when the possible negative consequences should the trust be disappointed are correspondingly more damaging than in other cases.

If we are inclined to accept premises (1) and (2), then the conclusion in (3) follows directly from them. But we face a problem with this: That preemptionism is false is evidenced by considerations that justify a degree of epistemic distrust. At the same time, (3) claims that we are unjustified in distrusting a BBA. Can we reconcile these opposing beliefs?

Healthy Mistrust: Between Preemption and Mistrust

I think there is a way to do this. For the considerations that speak against preemptionism can be used to describe more precisely what exactly the epistemic approach to BBAs should look like in order, on the one hand, not to distrust blindly (demand of (3)) and, on the other hand, not to trust blindly (demand of the critique of preemptionism).

First of all, the reference to the epistemic situation in which patients find themselves when they are confronted with the result of a BBA shows that we have to keep in mind the situation in which the diagnosis is communicated. If there is no prior opinion, for example, from other tests, then the acceptance of the result obtained by the BBA is in principle less of a problem than in other cases. Of course, there could be unease—and this could be justified—about the data basis of the algorithm or its reliability. But if it is not proven that the algorithm can reliably detect diseases, then this fundamentally calls into question whether it is an EA at all. In other words, the question of whether we can and should trust a BBA as an EA does not arise here at all. If, on the other hand, its epistemic superiority is proven, for example, through licensing, audits, and the like, then the distrust could be dispelled by pointing out how we also approach other EAs with trust in situations in which we have no prior conviction. This comes close to the argument often heard in medicine, according to which he who heals is right.

If there are prior opinions, then the adequate level of justification depends on whether the opinions are consistent with the result of the BBA. If, for example, there is a disagreement with other tests carried out independently of the algorithm—that is, if the result of the algorithm diagnoses something different from what other tests have shown or what the doctor says—then it becomes clear from the critique of preemptionism that it is necessary to aggregate our reasons and weigh them against each other. Preemption is ruled out if the criticism above is taken seriously, but other methods are available. One way is to compare the authority of different diagnostic systems or tests. The novice is thus faced with the problem of deciding which one of two experts to trust (*novice-two-expert problem*)—leaving aside for the moment that there is a difference between an expert and an EA.[25] Alvin Goldman identifies five sources of evidence that can be used to make an assessment of the reliability of the presumed expert, which, in another article, I subsequently use as a basis for considering a BBA as an expert:

"(1) it can evaluate the arguments made by the experts, (2) it can use the additional agreement of other experts, (3) it can use meta-experts' opinions about the expert in question, (4) it can use evidence about the expert's interests and biases to assess trustworthiness, and (5) it can use the expert's past track record of truth-seeking."[26] All five "sources of evidence" are also available in the case of algorithms. (1) is *by definition* not directly accessible, but there may be beliefs that give us a reason to believe that the expert's beliefs are reliable (Goldman calls this indirect justification). The success of algorithms would be one such indirect belief. Sources of evidence (2) and (3) are present, for example, when algorithms have received certain quality awards or have successfully passed audits. (4) is also relevant, for example, if it is disclosed (or not) what the data basis consists of with which the algorithm was trained and/or operates. And under the source of evidence (5), the success of an algorithm in other settings or in relation to other diseases can be subsumed. The same procedures can now be used to assess the expert or EA who conflicts with the BBA, in order to ultimately arrive at adopting one of the diagnoses put forward as a conviction without (or possibly with) preemption.

If, in a third case, the BBA diagnosis is consistent with prior beliefs, there is nothing to prevent—or even some evidence in favor—of including the fact that the BBA as EA arrived at the corresponding belief in the pool of reasons that support the correlating belief (the diagnosis). Even if it is assumed—which is not necessarily the case with a machine learning (ML) algorithm—that the doctor and the BBA use the same reasons in the development of their beliefs (*double counting*), there is again some evidence that the

patient is epistemically better off than if she bases her belief on only one set of reasons. For the mere fact that *two* EAs arrive at one and the same opinion is another reason that speaks for the truth of the respective diagnosis.[27]

Secondly, the references to epistemic safeguards and the gradual attribution of trust provide starting points for a healthy distrust of BBA—a distrust, however, that the EA of BBA takes quite seriously. Should it be compatible with the idea of trust to still trust even if epistemic safeguards are not completely dismantled, then there is nothing to be said against attributing special weight to the outcome of a BBA, but at the same time not considering the outcome as the only relevant outcome. In this way, for example, a possible development could be countered that shifts diagnostics (and possibly even more) purely to algorithms. For even if these can reliably judge whether a patient is suffering from certain diseases, in any case a supplement or assessment by other bodies is required. This can, but need not, be a human being, but in any case there should be mechanisms that act as an epistemic shield. A doctor as such a shield can, on the one hand, base her diagnosis on further tests or also personal experience both as a doctor and with the patient. But she can also be expected to assess the reliability of the algorithm used. In the case of BBA, she does not know the arithmetic operations and the logic of the algorithm, because by definition she cannot do that, just as no one else can. But she knows whether it has been used successfully in other fields or elsewhere, how it behaves in cases similar to the patient's, whether it is regularly reviewed, and whether alternatives are available. If a BBA is embedded in such a setting, which many women doctors see as the future of BBA, then there is nothing to stop a basic distrust from being reduced and trust from being built up if the framework of the setting is set up as epistemic safeguards. From epistemological considerations, an argument can be made as to why this is necessary, even against any countervailing tendencies in medical practice and policy. This also puts a stop to a possible *credibility excess,* which sometimes overshoots an actually important and helpful goal—to improve diagnostics through BBA: Algorithms know a lot and are also a lot better than humans. But giving them sovereignty over medical decisions alone can lead to harm. This is especially the case when, as has happened in the past, algorithms have a discriminatory effect, which was only recognized after the algorithms became widespread.

Thirdly and finally, the reference to the *stakes that* determine the level of trust is helpful for adjusting trust in BBA. We saw that for "life-or-death" decisions, the bar for trusting someone (or something) is higher than for more mundane decisions with less "dramatic" consequences. Thus, the implication might be to adjust trust in algorithms (or indeed algorithmic decisions) depending on what is at stake—and possibly who is affected as a patient or uses the BBA.[28] This supports the call for a *value-flexible design* that builds patient values and preferences into the algorithm, but also into the doctor–patient algorithm relationship.

Summary

Mistrust against BBAs could be seen as caused by or even justified by a fear of preemption. This, at least, was the suggestion made in this paper. It is one of the first attempts to apply epistemological theory to the ethics of BBAs and ethics in general. However, the argument pursued in this article tried to show that epistemic mistrust is unjustified because preemption is probably not the best way to deal with EAs. But although this argument leads to epistemic trust, the reasons to do so recommend a certain level of mistrust. To solve this dilemma, we proposed further recommendations for the proper use of BBAs as EAs. One such recommendation is that whether preemption or mistrust is called for depends on the situation, that is, whether prior beliefs about the diagnosis, for example, exist, and if so, if they are consistent with the BBA's result. Another insight derived from the analysis is that we have an epistemological argument to keep certain epistemic safeguards in place. The crucial point is that these safeguards need not be human. All that matters is that someone or something keeps an eye on the effectiveness of the BBA and whether it works reliably. When commentators require human oversight over the use of BBAs, what they could mean is that rather than keeping humans in the loop, what counts is that epistemic safeguards be kept in the loop. And a (potentially incomprehensible) algorithm could fulfill that role as much as a human doctor could do. A final recommendation requires that the level of

trust (in the sense of epistemic safeguards) justifiedly put on BBAs depends on the stakes attached to the decision. A further consequence of this is that it is unwise to completely and indiscriminatingly accept or reject BBAs. What matters is that we take a close look at the situation in which BBAs are used and their results are communicated. The "how" matters more than the "that."

Of course, there are many more questions that directly follow from this. First of all, we would have to introduce a threshold of stakes under which safeguards are not required and over which they are. It is never an easy task to do so, but it is beyond the scope of this paper to provide such a threshold. Moreover, this could be a task for further investigation, also empirical ones, to determine when people are ready to accept BBAs (i.e., when an additional safeguard is not needed) and when they are not (i.e., when an additional safeguard is needed). Secondly, and more fundamentally, it is an open question whether preemption is really at the core of people's mistrust against BBAs. I have provided no empirical support for this, and one reason is that I am not aware of any study that examines this question. So there is future work to empirically support the basic thesis here. More dramatically, though, one could question the very hypothesis itself or at least qualify that hypothesis. As was said above, it might just be a demand sui generis that people require understanding or understandability from their epistemic counterparts. There would be no more explanation required if this were so—and no argument could ease the epistemic discomfort people have toward BBA. However, whether this is so or whether, as an alternative perspective, people's unease is open to rational argumentation where argumentative proof of justified trust along with a concrete prescription for the design of BBAs (in terms of both the algorithm design and the implementation design) leads people to lower their distrust, is hereby handed over to the further discussion of this very proposal.

Acknowledgements. This paper is based on research funded by the German Ministry of Education and Health (EPAMed, FKZ 01GP2207).

Competing interest. The author declares none.

Notes

1. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philosophy & Technology* 2021;**34**(2):349–71. doi:10.1007/s13347-019-00391-6; Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics* 2021;**47**(12):e3; Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019;**1**(5):206–15; Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicines* 2018;**15**(11):e1002689.
2. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 2016;**3**(1):1–12.
3. Domingos P. *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books; 2015.
4. But see also Robbins S. A misdirected principle with a catch: Explicability for AI. *Minds and Machines* 2019;**29**(4):495–514.
5. See note 1, Braun et al. 2021. See also Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 2021;**47**(5):329–35 and Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. *Big Data & Society* 2016;**3**(2):1–21.
6. See note 1, Bjerring, Busch 2021.
7. See note 5, Duran, Jongsma 2021:3.
8. See note 5, Duran, Jongsma 2021:2.
9. Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*, 7th ed. New York, Oxford: Oxford University Press; 2013.
10. See note 1, Bjerring, Busch 2021.

11. Hatherley JJ. Limits of trust in medical AI. *Journal of Medical Ethics* 2020;**46**(7):478; Kerasidou CX, Kerasidou A, Buscher M, Wilkinson S. Before and beyond trust: Reliance in medical AI. *Journal of Medical Ethics.* 2022;**48**(11):852–6.
12. See note 1, Braun et al. 2021.
13. We might call all those EAs to whom we refer as epistemic entities, that is, entities that transmit claims with the intention of eliciting beliefs in us. This could be a more inclusive term that frees us from the need to consider only persons, media, or collectives as EAs. As one anonymous reviewer has suggested, defining EAs with reference to persons, institutions, or collectives would imply that BBAs are also persons, institutions, or collectives. Even though one might consider BBAs as institutions, it is not my intention to specify what ontological status BBAs or EAs have, but rather emphasize the point that we can consider BBAs as EAs and we do so when we take them as being capable of providing us with claims or beliefs because and to the extent they are in an epistemically better position that we are.
14. Jäger C. Epistemic authority. In: Lackey J, McGlynn A, eds. *The Oxford Handbook of Social Epistemology*. Oxford: Oxford University Press; forthcoming.
15. Zagzebski LT. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford: Oxford University Press; 2003.
16. Fricker E. Testimony and epistemic autonomy. In: Lackey J, Sosa E, eds. *The Epistemology of Testimony*. Oxford: Oxford University Press; 2006:225–50.
17. McMyler B. Trust and authority. In: Simon J, ed. *The Routledge Handbook of Trust and Philosophy*. New York: Routledge; 2020:76–87.
18. Keren A. Trust and belief: A preemptive reasons account. *Synthese* 2014;**191**:2593–615.
19. See note 14, Jäger forthcoming:16
20. Constantin J, Grundmann T. Epistemic authority: Preemption through source sensitive defeat. *Synthese* 2020;**197**(9):4109–30.
21. Dormandy K. Epistemic authority: Preemption or proper basing? *Erkenntnis* 2018;**83**(4):773–91.
22. See note 14, Jäger forthcoming:15.
23. See note 14, Jäger forthcoming.
24. Popowicz DM. "Doctor knows best". On the epistemic authority of the medical practitioner. *Philosophy of Medicine* 2021;**2**(2):1–23.
25. Goldman AI. Experts: Which ones should you trust? *Philosophy and Phenomenological Research* 2001;**63**(1):85–110.
26. Wolkenstein A. Müssen Algorithmus-basierte Entscheidungen erklärbar sein? Über black box-Algorithmen und die Ethik von Überzeugungen in der Mensch-Maschine-Interaktion. In: Friedrich O, Seifert J, Schleidgen S, eds. *Mensch-Maschine-Interaktion: Konzeptionelle, soziale und ethische Implikationen neuer Mensch-Technik-Verhältnisse.* Paderborn: Brill mentis; 2022:312–31.
27. See note 14, Jäger forthcoming
28. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics* 2019;**45**(3):156.

Cite this article: Wolkenstein A (2024). Healthy Mistrust: Medical Black Box Algorithms, Epistemic Authority, and Preemptionism. *Cambridge Quarterly of Healthcare Ethics* 33: 370–379, doi:10.1017/S0963180123000646