Neuroethics (2023) 16:22
https://doi.org/10.1007/s12152-023-09527-0

ORIGINAL PAPER

Why Won't You Listen To Me? Predictive Neurotechnology and Epistemic Authority

Alessio Tacca · Frederic Gilbert

Received: 25 October 2022 / Accepted: 7 August 2023 / Published online: 24 August 2023

Abstract From epileptic seizures to depressive symptoms, predictive neurotechnologies are used for a large range of applications. In this article we focus on advisory devices; namely, predictive neurotechnology programmed to detect specific neural events (e.g., epileptic seizure) and advise users to take necessary steps to reduce or avoid the impact of the forecasted neuroevent. Receiving advise from a predictive device is not without ethical concerns. The problem with predictive neural devices, in particular advisory ones, is the risk of seeing one's autonomous choice supplanted by the predictions instead of being supplemented by it. For users, there is a potential shift from being assisted by the system to being over-dependent on the technology. In other terms, it introduces ethical issues associated with epistemic dependency. In this article, we examine the notion of epistemic authority in relation to predictive neurotechnologies. Section 1 of our article explores and defines the concept of epistemic authority. In section 2, we illustrate how predictive devices are best conceived of as epistemic authorities and we explore the subject-device epistemic relationship. In section 3, we spell out the risk of harms interconnected with epistemic deferral. We conclude by stressing a set of preliminary measures to prepare users for the authoritative nature of predictive devices.

Keywords Autonomy · Brain device · Epistemic authority · Predictive neurotechnology · Epistemic deference · Trust

Introduction

The development of predictive brain technology, including implantable neural devices, is surging. From epileptic seizures to depressive symptoms, predictive neurotechnology is used and developed for a large range of applications. Numerous companies have started extracting neural data in various fashions, especially by developing a personalised biomarker of targeting-specific symptoms with predictive neural device [1]. Allowing ongoing monitoring of cerebral activities aimed at identifying and predicting specific outcomes (e.g. seizures, depressive episodes, or compulsive behaviours, etc.) is very promising.

In this article, we are interested in advisory devices; namely, predictive neurotechnology programmed to detect specific neural events (e.g. epileptic seizure) which allows user to take necessary steps to reduce or avoid the impact of the forecasted neuroevent. These predictive technologies suggest enormous potential to assist users' decisions and

A. Tacca (✉) · F. Gilbert
Philosophy Program, School of Humanities, CALE,
University of Tasmania, 41, Private Bag, Sandy Bay,
Tasmania 7001, Australia
e-mail: Alessio.Tacca@utas.edu.au

F. Gilbert
e-mail: Frederic.Gilbert@utas.edu.au

capacities for self-determination. These novel generations of neurotechnology could, in theory, have a large range of clinical and non-clinical applications [2, 3]. They could contribute to increasing agential capacities, including cognitive ones such as reasoning, learning, decision making, information retrieval and analysis. They may also predict unwanted outcomes (i.e., depressive episodes, addictive habits, socially reprehensible conduct). However, the neurotechnologies have not reached such advanced development yet.

The questions which will be the focus of our article is why users may choose not to "listen" the device predictions, and whether there are risks associated with embodying predictions. Interestingly, in earlier studies, we reported that users of implantable predictive neural devices were not "listening" to their advisory system. For instance, a patient declared: "I wasn't trusting [the implantable brain-computer interfaces]…I just ignore it anyway" or "I just wanted it out of my head" [4, 5]. Analysis of the results demonstrated that the "non-listener" users were lacking a degree of trust in the system. This lack of trust emerged from initial predictions being inaccurate or missing targets, which led users to disengage with the system, including stopping feedback inputs into the machine, inducing a cascade of inaccurate identifications.

One way of explaining these users' disengaged reactions is to see this lack of trust as an indication that users were not considering Brain-computer interfaces (BCI) as an epistemic authority. We also observed the opposite, where full trust in the device led to some fusional and symbiotic relationship with the predictions [4, 6]. Fusing and merging with a predictive computer is not without ethical concerns. The problem with predictive neural devices, in particular advisory ones, is the risk of seeing one's autonomous choice supplanted by the predictions instead of being supplemented by it (i.e. the smart system in my head knows better than me; why shouldn't I always listen to it?). For users, there is a potential shift from being assisted by the system to being over-dependent on the technology (e.g. increase control to users' decision-making while concomitantly augmenting control on users' decision-making) [2]. In other terms, it introduces ethical issues associated with epistemic dependency, or the absence of epistemic independence. By becoming overly reliant or dependent on the prediction, users may lose their capability for self-determination.

These ethical issues associated with epistemic dependency call for an investigation into epistemic authority in relation to predictive neurotechnologies. Sect. "Introduction" of our article explores and defines the concept of epistemic authority. In Sect. "Epistemic authority", we illustrate how predictive devices are best conceived as epistemic authorities, and we explore the subject-device epistemic relationship. In Sect. "Predictive devices are epistemic authorities", we spell out the risk of harms interconnected with epistemic deferral. We conclude by recommending a set of measures to prepare users for the authoritative nature of predictive devices.

Epistemic Authority

Predictive devices are used primarily because of their high level of accuracy and reliability. They are better at predicting targeted events than any human, including the user herself. In loose terms we could say that the device *knows* better than humans when a certain event is about to occur or is very likely to occur. In more technical terms, the predictive device is in a better epistemic position than the subject (S) when it comes to formulate propositions (p) in a specific domain (D). Epistemologists call those who are in better epistemic position *epistemic authorities*. We propose the following definition of epistemic authority:

> A is an epistemic authority for S with respect to propositions (p) in a specific domain D, just when S thinks that by relying on A's propositions, S is in a better epistemic position than they would be if they did not rely on A's claims.

Some aspects of the definition need to be explained.

S Thinks Our definition requires that S *thinks* that they will end up in a better epistemic position, should they rely on A's propositions. It is not at all sure that they will be. A could well make a wrong or inaccurate claim, but our definition of epistemic authority does not require the truth of the authority's propositions. If A possessed a solid record of true statements in D, then surely S would be more likely to consider them an epistemic authority in that domain D – but this is not, strictly speaking, necessary. S could still not retain A an epistemic authority even if A had an impressive record of true statements.

Relying By "relying on A's propositions in D" we mean that S forms beliefs and likely ends up acting on them, in virtue of having assumed A's propositions as true. Note that S must not necessarily know how A came to know p in D. Once again, according to our scientific standards, knowing that A produced p thanks to accurate scientific research, with the aid of an algorithm having monitored S's brain for several years, as opposed to having been told p by an algorithm set to produce random outcomes, increases the likelihood of S considering them an epistemic authority. But again, that is not necessarily the case. S might still be more inclined to trust someone who has been told things in dreams than professional researchers. As strange as it sounds, that could be the case, and our definition tries to capture this possibility.

Better Epistemic Position The expression we chose is deliberately unspecific. The reason is that we did not want to stick to one specific framework for characterising what exactly is the epistemic advantage of S relying on A's propositions. This advantage can be framed in several ways. S could, for instance, be said to be aiming at accuracy as their primary epistemic goal. In this case, they would rely on A's propositions were they considering A to have a higher expected accuracy in respect to propositions in D.[1] However, the concept of being in a better epistemic position could also be framed in terms of justification. Accordingly, S would be more *justified* in believing some propositions in D if these propositions were relying on evidence coming from A, compared to other evidence they could get without A.[2]

The definition of EA we presented is fundamentally an internalist one. The debate between internalists and externalists concerns the nature and source of justification. Internalists hold that factors relevant to justification must in some way be reflected in the agent's beliefs, whereas externalists deny this. In our definition S's belief that a certain source of information (person or device) is an EA is necessary and sufficient to *make* that source an EA *for S*. This is indeed an internalist position. However, we also believe that there are parameters for evaluating sources of justification (or even knowledge) that communities maintain to be more reliable than others. For example, in a veritist community, where truth is held as the primary epistemic goal, truth conduciveness and positive track-record of true outcomes are considered important and perhaps essential properties that A has to possess in order to be an EA. However, even within this community, a single subject S could still think that trusting tea leaves instead of doctors will put them in a better epistemic position, whether that means possessing a clearer understanding of their health condition or being more likely to "get things right" about what cure to undergo or whatever other epistemic goal S may have. So, to an external viewer, the member of the veritist community S could appear as to possess (knowingly or not) justification for their beliefs about their health because, say, they have been to reliable doctors. But in reality S may still not consider the same reliable doctors as epistemic authorities. Instead, they think they are put in a better epistemic position if they follow tea leaves "advise", which may or may not be in line with the reliable doctors opinion.

It is possible to amend our definition of EA to turn it into an externalist one:

[1] This solution is proposed by Bokros, who defines EA within the framework of accuracy first epistemology: "A is an epistemic authority for S with respect to p iff S judges A to have a higher expected accuracy with respect to p than S takes herself to have independently of following A's authority" [7]. In many ways, our definition is a development of Brokos'.

[2] In turn, possessing solid and reliable justification for p, would bring S closer to having actual knowledge that p. The emphasis on justification is a feature of much of the early discussion on externalist reliabilism. Note that here we are only discussing S's justification for p. Two other important questions worth asking are: 1) whether S is justified in regarding A as an EA and 2) how S acquires this justification. The sources of S's justification for such belief could be of different nature, they could be internal, external, from testimony, direct experience and so on. A proper discussion of this topic would deserve a separate article and constitutes perhaps one of the most interesting developments of this research, from an epistemological perspective. What we want to focus on in this article, however, is defining what an EA is for S in respect to some specific proposition p in domain D. This is an issue that, in our opinion, precedes a discussion on the sources of S's justification for believing that A is an EA. Thank you for reviewer 1 for pointing at this element.

> **Ext. EA DEF**. A is an epistemic authority for S with respect to propositions (p) in a specific domain D, just when by relying on A's propositions, S is in a better epistemic position than they would be if they did not rely on A's claims.

However, this externalist characterisation of EA implies the existence and identification of external means of assessing whether S is in fact in a better epistemic position by following A's advice or not. This would in turn imply fixing a set of rules or standard for when S is *de facto* in a better epistemic position regardless of what S thinks. This sounds problematic. Surely, an externalist may appeal to the presence of some reliable casual processes in every determination of EA, but this would imply arguing that the person who decided to trust tea leaves as EA instead of doctors shared and applied the same casual process of the person who did the other way around.

Furthermore, even conceding to the externalist that there are identifiable trustworthy processes to externally determine that A either is or is not an EA for S independently of what S thinks, S is still free to choose to trust B instead of the externally-recognised epistemic authority A because they *fail* to recognise A as the EA. For example, S could put their trust in tea leaves instead of doctors although doctors are externally recognised as EA.

So, given the problems presented by an externalist view of EA, we believe that an internalist definition better captures the dynamic of trust and action upon the evidence produced by what S *thinks* is an EA. We remain open to the possibility of revision of our definition in an externalist direction were the problems we presented in relation to those views be resolved.

Bokros supports a hybrid version via the concept of epistemic superiority. They define epistemic superiority as follows: "A is epistemically superior to S with respect to p iff A has a higher expected accuracy with respect to p than does S" [7]. Then they claim that "epistemic superiority is objectively defined whereas authority is a subjective relation; it is dependent on whether S makes the judgment [we said "think"] that A is epistemically superior to herself" (ibid). Obviously then they run into the problem of defining (objectively) expected accuracy, which seems rather slippery given that something "expected" is inherently open to misjudgements and error.

After the predictive device passes the optimal calibration phase, and synchronises with brain activity and symptoms, we argue that it acts as an epistemic authority for the user. This is true even though the machine is not necessarily understood as an *expert*, who is normally a figure associated with great knowledge and therefore an authority in a specific domain. Goldman [8] defines an expert as someone who "must possess a substantial body of truths in the target domain [and] a capacity or disposition to deploy or exploit this fund of information to form beliefs in true answers to new questions that may be posed in the domain" (p. 91). Although the predictive device may satisfy the first of Goldman's conditions, namely possessing a substantial body of truths in a specific domain, the fulfillment of the second condition is doubtful. It is in fact problematic to attribute to a machine any capacity to form beliefs or have *dispositions* at all. So, unless we reject Goldman's definition – which we do not find useful to do – the predictive device is not to be considered an expert.

Considering our definition of epistemic authority, it is reasonable to conclude that if one qualifies as an expert by Goldman's standards, that puts them is a strong position to also be an epistemic authority in that domain. Another way of putting this is to say that experts are more likely to possess properties that are required to be an epistemic authority.[3] However, despite the close relation between expertise and epistemic authority, we agree with Bokros [7] that being an expert is not necessary nor sufficient for being an epistemic authority.[4] To prove that predictive devices

[3] There is no clear agreement on what these properties are exactly, but they include reliability, accuracy, success rate etc.

[4] An externalist, according to whom it is possible to define epistemic authorities for S independently of whether or not S thinks they are, may identify being an expert as an either sufficient or necessary condition for being an EA for S. We think being an expert is not a sufficient condition for being an EA because that would imply that all experts are EA. But that is not possible since, according to the internalist view proposed, we still need S's belief that A is an EA to make it an EA *for S*. In respect to an expert being necessarily an EA, we find the view problematic primarily because S could still consider tea leaves and soothsayers as EA while admitting that neither are experts. Note that we do not advocate for the unimportance of expertise. On the contrary, we do believe that experts are good candidates for being considered EA. What we reject is the idea that S plays no role in determining what is an EA *for them*. Again, externalist standards to characterise ideal EAs according to scientific standards, causal reasoning, good track-record of epistemic success etc. may be worked out, but that does not take away the fact that if S does not think A is an EA *for them*, then A is not an EA *for them*. Thanks to both reviewers for having asked to clarify this aspect.

can be epistemic authorities, we first argue that not only experts are epistemic authorities, and then show how predictive devices fulfill the standards set by our definition of epistemic authority.

Bokros considers a scenario in which “a grandmother could figure as epistemic authority for her young grandson on the topic of how fish breathe, despite only having elementary knowledge of zoology, because her grandson has no knowledge of the topic at all” (p. 12050). In this scenario, however, we could still argue that the young grandson *considers* his grandmother an expert. After all, he has not yet sharpened his cognitive tools to rationally evaluate levels of expertise in unknown domains. Furthermore, even if faced with a zoologist who contradicts his grandmother, the young grandson would most likely still refrain from trusting the strangers. Were the grandson capable of determining who the expert was between the zoologist and his grandmother, and were he a veritist aiming at knowing the truth about how fish breathe, then he would consider the zoologist an epistemic authority. There are, however, other cases in which the subject treats their source as an epistemic authority despite knowing that the source is not an expert.

Consider the case of a dog who is always correct at predicting his master epileptic seizures [9]. If questioned on how he manages to get things right all the time, the dog would not produce any reasonable explanation to justify his outcomes. A person might opt to trust the dog and even to consider it an epistemic authority, perhaps convinced by the 0% error record, but they would *know* and would unproblematically claim that the dog is not an expert in epilepsy. In fact, being an animal, the dog has no specific training in the domain in which he formulates his correct guesses. Still, that would not change the dog’s role as an epistemic authority for a subject S, who *decides* to act upon the dog’s predictions and is convinced that acting upon them would put them in a better position.

Establishing that experts are not necessarily epistemic authorities is relevant in at least two ways. First, because it leaves open the possibility for non-experts to act as epistemic authorities, including predictive devices, as we want to argue. Second, it shows that the vast literature on the epistemology of expertise, although it offers valuable insights, does not encompass all the epistemic issues we find in the subject-predictive device relationship. We need to at least adapt and expand the recent debate to better capture and explain the specifics of the relationship between human and predictive devices.

Predictive Devices are Epistemic Authorities

When it works accurately, the device produces predictions that effectively increase the chances of S obtaining the result they are after, say the prevention of a seizure. At an epistemic level, S may not increase their general understanding of the relevant underlying mechanisms. In other words, S may not know *why* the machine has produced the prediction it has produced. Consequently, S may not be able to use the machine processes as piece of evidence for their belief that, say, they are about to have a seizure. What they would use as evidence for their belief that they are about to have a seizure is *the device’s prediction itself*, meaning the fact that the device has made *that particular* prediction. The process that the machine has followed to arrive at its prediction may remain forever unknown to S, but that does not take anything away from the fact that S *thinks* they are *more justified* in believing that they are about to have a seizure, since the device has told them so. We will come back to ‘black box’ issues and ignorance of the devices’ processes later, but for the moment, the point we want to stress is that S’s ignorance of the device’s process that has led to a certain prediction does not mean that S is not in a better epistemic position after having followed the device’s advice. This fits with our definition of epistemic authority: if S thinks they are in a better epistemic position by following the device’s advice (A), that is enough for creating the conditions for A being an epistemic authority for S.

We have established that the predictive device is best characterised as an epistemic authority for the user. Now, to better understand the device-user epistemic relation to anticipate potential harms which might be induced, we need to focus on *how* the user engages with the machine and, in particular, with the machine’s outputs, namely predictions and recommendations. In the recent literature there are two main competing views in respect to how the subject does or should engage with an epistemic authority’s claims: the preemptive view, and the total evidence view.

Linda Zagzebski [10, 11] is the most prominent supporter of the preemptive view. She argues that “what is essential to authority is that it is a normative power that generates reasons for others to do or to believe something preemptively. […] A preemptive reason is a reason that replaces other reasons the subject has. […] The authority does give me a reason to believe or do something that replaces my other reasons relevant to the belief or act. The kind of reason authority gives me is what is essential to it" [11]. In the context of predictive neurotechnologies, this translates into a total reliance on the device predictions. The fact that the device is an epistemic authority means that it provides the subject with the only reason they need to believe the content of the prediction. In other words, the mere fact that the device (epistemic authority) says p is the one and only reason for S believing that p.[5]

In contrast to the preemptive view, the total evidence view sees the authority's claims as providing reasons for the S to believe p, but not pre-emptively. The authority's claims are just an addition to the other set of reasons S might have for p. As Dormandy [14] illustrates, if S based her beliefs only on the authority's reasons, S's other reasons for the same belief p would go unutilised. Consequently, this would put S in a weaker epistemic position, for she would have less justification than if she based her belief p on other reasons too (p. 774). Although prima facie the total evidence view seem to validate the subject more than the preemptive view by admitting their own evidence and letting it contribute to the decision-making process, in the context of predictive neurotechnologies, this can create problems. In fact, it may be dangerous for the user to validate their beliefs with supporting evidence coming from places other than the predictive device, where this evidence conflicts with the device's predictions. Validating and supporting this clash of beliefs may induce the user to think that the device predictions count exactly as much as any other evidence. Yet, given the high level of accuracy of predictive devices in some medical contexts we know that this is not true.

The employment of a preemptive or a total evidence framework affects our understanding and interpretation of the user-device relation. In medical cases, it also affects clinicians' provisions and recommendations. A supporter of the preemptive view may be more inclined to instruct their patient to follow the device predictions and disregard other reasons for either the same or an alternative course of actions. This is motivated by the fact that the device is an epistemic authority and the epistemic authority's reasons outweigh any other reasons. A supporter of the total evidence view, in contrast, may be more inclined to instruct their patient to take the device predictions not as necessarily true, but to combine them with other reasons for or against the device's recommended course of action. Both the preemptive view and the total evidence view have some risks of harm in common. Let us look at those risks.

The Risks of Epistemic Deference and Potential Harms

Relying on an epistemic authority comes at the cost of some risks, which Fricker calls the risks of deference [15]. See Table 1.

Loss of Trust

While Fricker also identifies the risk of deliberate deception, we assume that predictive devices have no dispositions, character, volition, and so on. We can therefore exclude instances in which a device deliberately and autonomously deceives the user. A sophisticated AI coupled with the predictive device so cleverly designed as to be capable of producing "white lies" is conceivable. However, to the best of our knowledge, for the moment no such AIs are employed in the devices we are considering. If the situation changes in the future, then we would reassess this point, but for the moment we can put aside the issue of deliberate deception.

Still, the machine may be wrong. Its prediction may turn out to be false or inaccurate or misleading or unclear. The main risk associated with such events is the loss of trust from the user. In cases in which the device is implanted inside the user's body, this loss of

[5] Zagzebski's view has attracted several criticisms. In particular, there is a concern that grounding beliefs exclusively on an epistemic authority's claims ultimately threatens critical thinking. Constantin and Grundmann [12] consider this objection and propose an amended version of the preemptive view which, according to them, avoids it. Another concern is legal in nature. The preemptive view could exacerbate responsibility gaps which, as some scholars have noticed [13], are a concern for many BCI technologies.

Table 1 We have applied Fricker's three main categories of deferential risks to the specifics of the user-predictive device relationship and illustrated how these could translate into potential harms for the user

Categories of deferential risks	Deferential risks translate into predictive neurotechnology harms
A Deception and error (deliberate or accidental)	A Loss of Trust: Wrong predictions. May lead to loss of trust, desire to interrupt treatment. Device is no longer considered an epistemic authority
B Dependence	B Dependence and loss of autonomy: Lessened ability to manage portions of one's life. E.g. inability to recognise symptoms, over-reliance on machine outputs. Leads to lack of autonomy, or perceived lack of autonomy. Cognitive outsourcing/offloading. Addiction to the machine
C Inability to "police" one's system of beliefs	C Blind trust and black box: 'I don't know why, but the machine said I must do x'. Leads to 'blind trust' and acritical approach. Black box issues

trust may have substantial psychological and practical impacts, from the desire to remove the device (often requiring an operation), to an extended loss of trust in the treating team and their advises, with obvious consequences on the patient's health and wellbeing.

Potential issues of loss of trust could be mitigated or even prevented with targeted, continuous, and holistic education. It is not enough to provide a medical explanation of how and why the device is the best treatment option. Potential end users must be also introduced to the devices operating principles. We address this aspect in our recommendations below.

Dependence and Loss of Autonomy

Zagzebski [11] defines autonomy as "the executive self's management of itself, or what is usually called self-governance" (p. 230). It requires one being conscious of one's self-consciousness and ultimately aims at having "the right relation to a world outside my mind". This is obtained by "exercising my faculties in the best way I can to make the outputs of those faculties fit their objects—to make my beliefs true, my desires of the desirable, my emotions appropriate to their intentional objects" (ibid).

In Zagzebski's characterisation of autonomy, there is a strong connection between being autonomous and being right. In fact, it seems that it is only by being right in my claims and beliefs about the world that I can obtain accordance with the external world. If I were free to act and believe whatever I wanted, but constantly got things wrong, by this characterisation of autonomy, I would not be autonomous. I would only be free to do things wrong. Autonomy requires a degree of rationality and rationality aims at getting things right. Whether it succeeds or not is a separate issue, but there is nonetheless a strong connection between rationality and truth in that one constantly seeks the other. So, if autonomy is not to be conceived solely as freedom to do and believe whatever we want, but as a more complex series of self-governed actions based on rational mental processes which aim at accordance with the external world, then we can see how relying exclusively on ourselves when we engage with the world at any epistemological level (e.g. by trying to predict something or seeking evidence to support or reject the truth of a certain belief we have etc.) might be more *autonomous* if we do rely on some external epistemic authority.

Zagzebski calls "epistemic self-reliance" the attitude of relying exclusively on justification coming from within ourselves to formulate judgements about the world. The idea of epistemic self-reliance being at the core of any rational investigation is extremely hard to give up, since it has been considered pivotal in the history of philosophy (Locke, Descartes) and arguably sits at the basis of some fundamental theories and views in western society (individualism, liberalism, autonomy, and moral responsibility, to name a few).[6] As we have seen, however, epistemic self-reliance not only *can* lead to errors, but in many cases is more likely to do so. These are cases in which the subject has no

[6] Epistemic self-reliance has only been challenged relatively recently in epistemology, with discussions on reliabilism, the epistemic value of testimony and externalism in general.

expertise and would greatly increase their chances of producing accurate beliefs if they were to rely on some external epistemic authority. In turn, given how we defined autonomy previously, that would then give them more autonomy.

A similar argument against the loss of autonomy when relying on an epistemic authority other than oneself, comes from Fricker [15]. She argues for the rationality of accepting an external epistemic source, and even for allowing it to override our prior opinion on a specific matter, if we know that this source “is strongly placed epistemically, and better placed than oneself, regarding the matter in question”. It is crucial, however, that our trust “is given not blindly and universally, but discriminatingly. By trusting only cannily, and with good grounds, we can do much to retain epistemic self-governance” (p. 239). The discussion of the standards required to determine in which cases we have good grounds for trusting an external epistemic source falls beyond the scope of this article. However, the relevant point is that we have, in theory, a possibility or determining them.

In the specific case of predictive devices, the patient is always in control of the actions that they deliberately decide to perform, whether they are in accordance with the device’s recommendations or not. For Mele [16], in fact, the best way to understand autonomy is in terms of control. According to this account, autonomous decisions require that an individual exercises control over their decisions. In that respect, autonomy encompasses self-control or being a self-controlled individual. However, identifying or endorsing a sort of control is not sufficient to demonstrate that a user’s decision counts as having been made autonomously. As pointed out by many scholars [17, 18], the threat to autonomy is not external but rather internal to the user’s decision-making process. If it was external, any predictive neural device would be a menace to autonomy (e.g. external controlling influences). The potential threat to autonomy likely resides where there is no internal distinction by the user (e.g. internal controlling influences). Some users may be at risk of over-reliance on advisory devices. It has been argued elsewhere that this translates into *decisional vulnerability* when patients are faced with forming a decision to follow the device’s information [2, 4]. Decisional vulnerability occurs in a context of epistemic dependence, in particular when patients outsource their deliberative capacities to device instructions despite the presence of immediate evidence advising otherwise (e.g. device providing a false sense of security, user putting themselves in risky circumstances). In extreme cases, these internal controlling influences may be manifested into addictive behaviour (e.g. rendering a user incapable of making a decision without systematically consulting their devices; preferring the advisory system over recommending medication, etc.).[7]

Blind Trust and Black Box

Even if presented with an accurate description of the intricate passages that led the device to formulate the prediction it did, the user would most likely not understand it. First, most likely the user lacks training in AI or machine learning, and second, even if they did have that specific knowledge, they may still not be able to understand *why* the machine has processed data in *that* particular way, to form *that* particular prediction. The literature on black box models in AI machine learning shows that a lack of detailed understanding of all processes and steps taken by a sophisticated AI to produce a certain outcome raises issues not only for the experts employing these devices – e.g. a medical treatment team – but for AI experts alike. Wadden defines black box issues in healthcare AI as occurring “whenever the reasons why an AI decision-maker has arrived at its decision are not currently understandable to the patient or those involved in the patient’s care because the system itself is not understandable to either of these agents” [21]. Black box issues have the potential to generate loss of trust, and thereby impact the success of treatment. More research is still needed in the understanding of the epistemological and psychological impacts of the potential employment of black box system in predictive neurodevices.

Blind trust is another issue deriving from the subject’s incapacity to control their system of belief. It is typically generated by the user over-relying on device predictions. Imagine a user implanted with a device for long-term treatment resistant depression,

[7] In this section, autonomy issues have been discussed with specific reference to predictive neurodevices. Other scholars have investigated issues of autonomy and BCI technology more broadly [19, 20].

who then loses her beloved mother. She is then advised by the system to take anti-depressant medication. If the user trusted the device blindly, she would take the drug. However, being sad at the loss of her mother is likely the most appropriate behaviour. Not grieving may cause even bigger problems than being depressed after her passing [22]. This is a typical case in which blind trust leads to foreseeable non-beneficial outcomes. However, similar to the case of potential autonomy issues, whether blind trust is a problem for the predictive device user depends on how we define blind trust and the epistemic role we assign to testimony. Traditionally, epistemologists insist on the premise that knowledge relies on evidence, not trust. However, at least since the early Nineties [23], the literature on trust in epistemology has shown that knowers are not completely independent and self-reliant. If they were, they would often fail to possess the best evidence for their beliefs and they would end up knowing far less than we ordinarily claim to know. Forming beliefs based on someone else's claim naturally involves trust, and with it comes a certain degree of *blindness* in respect to our trustee's reasons. Yet this is most of the time unproblematic. Ignorance of the steps that has led our source to produce the claim they did is a feature of almost all knowledge acquired by testimony. When we rely on someone else's expertise, we do not necessarily care about knowing all steps of their reasoning, all pieces of information they collected and put together to produce the claim they did. We just trust that what they say is more likely to be true compared to what we could have said in that domain, without relying on their opinion.

The issue for the subject is primarily that of identifying reliable and trustworthy sources, more than working out an accurate reconstruction of the sources' reasons, so that the subject can turn them into own's evidence. This latter is often unnecessary, convoluted and time consuming. Once we establish that a source is trustworthy in a domain D, we rely on it without going any further in the investigation of how they acquired their knowledge in D.

If we apply this reasoning to predictive devices, knowledge of the processes that led to the device's predictions and recommendations is not necessary, provided that the user *trusts* the device. Trustworthiness and reliability are difficult to frame at a theoretical level. Trust cannot be analysed as a purely epistemological element. If S considers a certain source B trustworthy, this often means not only that S beleives that B 'knows better' in a certain domain, but also that it would be better to trust B. There is a fundamentally moral component to trust. For instance, S may need to trust that B would not use their knowledge to harm or deceive them. So, for B to be trustworthy to S, B needs to be both epistemically and morally reliable according to S' standards. The degrees of epistemic and moral reliability, and the standards employed in their evaluation, depend on each circumstance and subject.

Conclusion: You Should Listen To Me But…

Predictive neural devices are powerful algorithmic-based systems which provide personalised advice to users. They present uniquely individualised risks of physiological and psychological harms, which are proportionally intertwined with concerns linked to epistemology. The current article on predictive neurotechnologies raises a crucial epistemological question: are predictive neurotechnologies an epistemic authority? While the answer to this question may not always be clear, it should always have a direct bearing on conclusions concerning what should (or should not) be ethically prescribed for users. Underestimating the epistemic challenges associated with predictive BCI may lure prospective patients into thinking the experimental trial is without risks of actual harms. Furthermore, despite notable exceptions [24], epistemological issues associated with BCI technologies are relatively under investigated compared to the ethical issues associated with the same technologies [25, 26].

As discussed in Sect. "Dependence and loss of autonomy", although relying on an epistemic authority does not necessarily threaten the subject's autonomy, there are still some risks. Interpreting the epistemic relation between predictive devices and their users should characterise the way candidate users are instructed and prepared for treatment with predictive neurotechnology devices. Obviously, we believe that every user-device relation is different. Accordingly, explaining how a patient evaluates the device's epistemic authority and engages with it is not an easy task. Given the variety of devices as well as the differences in character traits, hopes, desires and so on of patients, it is difficult to pin down exactly a standard procedure for patients to follow when dealing with their predictive devices.

As we have argued in Sect. "Epistemic authority", the device does not need to be an expert to be conceived as epistemic authority by the user. As such, normative practices can be derived from the device's recommendations. One point deserving further development is whether the predictive device user is required to believe p in order to act upon it. In our view, if S maintains A to be an epistemic authority in D, S maintains automatically that A's predictions in D are more likely to be true. Therefore, it would be more epistemically beneficial to act on them. Our view hinges on the belief that there is an intrinsic normative value in epistemology which, at its minimum, holds that if one does not believe that p is true, then they should not act on it. Conversely, they should, if they believed p was true. Whether S effectively ends up acting or not acting according to p depends on many social, psychological, situational and other factors. But if the goal is that of achieving a beneficial relationship between predictive devices and their users, especially in medical contexts, it is paramount that we care about consistency between what patients believe is true or best for them, and their actions. It is not enough to be content with the users *just* doing what the device told them.[8]

As a standard ethical measure, we highly recommend that clinicians and medical teams articulate clear protocols to inform and guide users about the potential inaccuracies and algorithmic misalignments that may arise when using predictive devices. During the calibration phase, clinicians must ensure that users are adequately prepared to acknowledge this possibility and have a robust understanding of how the predictive device operates to prevent blind trust issues, as described in Sect. "Blind trust and black box". Moreover, it is crucial to ensure that users are aware that relying solely on the device as an epistemic authority may have ethical implications that need to be considered. Nevertheless, clinicians should remind users that utilizing the device as an epistemic authority can significantly increase the chances of receiving effective treatment.

The debate over whether predictive neurotechnologies should be considered epistemic authorities has direct ethical implications, as the conclusions drawn from these technologies can have significant consequences for patients' lives. Therefore, clinicians should be aware of some of these debates and engage with them critically so that they can make informed recommendations to patients and address their concerns about the reliability and validity of predictive neurotechnologies. The task of involving clinicians in the debate also falls on the shoulders of scholars in epistemology. Philosophers could benefit greatly from having clinicians contribute to theoretical discussions. There are many models and examples where neuroethicists have clinicians as co-authors, which increases the scope, impact and magnitude of any ethical conclusions and applications of their studies. It might be time for philosopher epistemologists to follow a similar co-authorship model, certainly when theory might have significant effects for patients' lives. As a result, clinicians could more easily endorse these epistemological conclusions through their medical prescriptions with increased awareness, and better address potential patients' issues and concerns.

In conclusion, clinicians have an important role to play in the ethical and epistemological debates surrounding predictive neurotechnologies. By engaging with these debates and considering the ethical implications of their medical recommendations, clinicians can help ensure that the use of these technologies is guided by sound ethical principles and a critical assessment of their epistemic value.

Acknowledgements We are grateful to Neuroethics anonymous reviewers for their valuable comments.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflict of Interest None.

[8] The primary interests of this paper are epistemological and ethical in nature. Buller [27–29] provides a richer and more focused discussion on BCI and theories of action, whereas other scholars focus on the consequences of BCI for concepts of agency [30].

References

1. Haeusermann, T., C.R. Lechner, K.C. Fong, A. Bernstein Sideman, A. Jaworska, W. Chiong, et al. 2021. Closed-Loop Neuromodulation and Self-Perception in Clinical Treatment of Refractory Epilepsy. *AJOB Neuroscience* 2: 1–13.
2. Gilbert, F. 2015. A Threat to Autonomy? The Intrusion of Predictive Brain Implants. *AJOB Neuroscience* 6 (4): 4–11.
3. Miletic, T., and F. Gilbert. 2020. Does AI Brain Implant Compromise Agency? Examining Potential Harms of Brain-Computer Interfaces on Self-Determination. Ed Gouveia S.S., In: *Artificial Intelligence and Information: A Multidisciplinary Perspective*. Vernon Press. Pages 253-272. ISBN: 978-1-62273-872-4
4. Gilbert, F., M. Cook, T. O'Brien, and J. Illes. 2018. Embodiment and Estrangement: Results from a First-in-Human "Intelligent BCI" Trial. *Science and Engineering Ethics* 25 (1): 83–96.
5. Gilbert, F., T. O'Brien, and M. Cook. 2018. The Effects of Closed-Loop Brain Implants on Autonomy and Deliberation: What are the Risks of Being Kept in the Loop? *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees* 27 (2): 316–325.
6. Gilbert, F., Ienca, M., Cook, M. 2023. How I became myself after merging with a computer: Does human-machine symbiosis raise human rights issues?. *Brain Stimulation* 16 (3): 783–789 https://doi.org/10.1016/j.brs.2023.04.016
7. Bokros, S.E. 2021. A deference model of epistemic authority. *Synthese* 198 (12): 12041–12069.
8. Goldman, A.I. 2001. Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research* 63 (1): 85–110.
9. Powell, N.A., A. Ruffell, and G. Arnott. 2021. The Untrained Response of Pet Dogs to Human Epileptic Seizures. *Animals: An Open Access Journal from MDPI* 11 (8): 2267.
10. Zagzebski, L. 2013. A Defense of Epistemic Authority. *Res Philosophica* 90 (2): 293–306.
11. Zagzebski, L.T. 2012. *Epistemic authority: a theory of trust, authority, and autonomy in belief*, 279. Oxford ; New York: Oxford University Press.
12. Constantin, J., and T. Grundmann. 2020. Epistemic authority: Preemption through source sensitive defeat. *Synthese* 197 (9): 4109–4130.
13. Bublitz, C., A. Wolkenstein, R.J. Jox, and O. Friedrich. 2019. Legal liabilities of BCI-users: Responsibility gaps at the intersection of mind and machine? *International Journal of Law and Psychiatry* 1 (65): 101399.
14. Dormandy, K. 2018. Epistemic Authority: Preemption or Proper Basing? *Erkenntnis* 83 (4): 773–791.
15. Fricker, E. 2006. Testimony and Epistemic Autonomy. In *The Epistemology of Testimony*, ed. J. Lackey and E. Sosa, 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276011.003.0011. (cited 2022 Aug 31).
16. Mele, A.R. 1995. *Autonomous Agents: From Self Control to Autonomy*. Oxford University Press.
17. Glannon, W. 2014. Neuromodulation, agency and autonomy. *Brain Topography* 27 (1): 46–54.
18. Müller, S., and H. Walter. 2010. Reviewing autonomy: Implications of the neurosciences and the free will debate for the principle of respect for the patient's autonomy. *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees* 19 (2): 205–217.
19. Wolkenstein, A., and O. Friedrich. 2021. Brain-Computer Interfaces: Current and Future Investigations in the Philosophy and Politics of Neurotechnology. In *Clinical Neurotechnology meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, ed. O. Friedrich, A. Wolkenstein, C. Bublitz, R.J. Jox, and E. Racine, Cham: Springer International Publishing. [cited 2022 Oct 21]. p. 69–80. (Advances in Neuroethics). https://doi.org/10.1007/978-3-030-64590-8_6.
20. Friedrich, O., E. Racine, S. Steinert, J. Pömsl, and R. Jox. 2021. An Analysis of the Impact of Brain-Computer Interfaces on Autonomy. *Neuroethics* 1: 14.
21. Wadden, J.J. 2021. Defining the undefinable: the black box problem in healthcare artificial intelligence. Journal of Medical Ethics. [cited 2022 Sep 22]; Available from: https://jme.bmj.com/content/early/2021/07/20/medethics-2021-107529. Accessed 17 Sep 2022.
22. Klein, E., S. Goering, J. Gagne, C.V. Shea, R. Franklin, S. Zorowitz, et al. 2016. Brain-computer interface-based control of closed-loop brain stimulation: Attitudes and ethical considerations. *Brain-Comput Interfaces.* 3 (3): 140–148.
23. Hardwig, J. 1991. The Role of Trust in Knowledge. *The Journal of Philosophy* 88 (12): 693–708.
24. Schleidgen, S., O. Friedrich, and A. Wolkenstein. 2022. How intelligent neurotechnology can be epistemically unjust. An exploration into the ethics of algorithms. *Review of Social Economy* 80 (1): 106–26.
25. O. Friedrich, and A. Wolkenstein. 2021. Introduction: Ethical Issues of Neurotechnologies and Artificial Intelligence. In *Clinical Neurotechnology meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, ed. O. Friedrich, A. Wolkenstein, C. Bublitz, R.J. Jox and E. Racine, 1–9. Cham: Springer International Publishing; [cited 2022 Oct 21]. (Advances in Neuroethics). https://doi.org/10.1007/978-3-030-64590-8_1.
26. Wolkenstein, A., R.J. Jox, and O. Friedrich. 2018. Brain-Computer Interfaces: Lessons to Be Learned from the Ethics of Algorithms. *Cambridge Quarterly of Healthcare Ethics* 27 (4): 635–646.
27. Buller, T. 2021. Brain-Computer Interfaces and the Translation of Thought into Action. *Neuroethics* 14 (2): 155–165.
28. T. Buller. 2021. Actions, Agents, and Interfaces. In: *Clinical Neurotechnology meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, ed. O. Friedrich, A. Wolkenstein, C. Bublitz, R.J. Jox, E. Racine, 11–23. Cham: Springer International Publishing; [cited 2022 Oct 21]. (Advances in Neuroethics). https://doi.org/10.1007/978-3-030-64590-8_2.

29. Buller, T. 2020. How to Do Things with BCIs. *AJOB Neuroscience* 11 (1): 70–72.
30. Steinert, S., C. Bublitz, R. Jox, and O. Friedrich. 2019. Doing Things with Thoughts: Brain-Computer Interfaces and Disembodied Agency. *Philosophy & Technology* 32 (3): 457–482.

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