Medical AI, inductive risk and the communication of uncertainty: the case of disorders of consciousness

Jonathan Birch

Centre for Philosophy of Natural and Social Science, LSE, London, UK

Correspondence to
Professor Jonathan Birch, Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, WC2A 2AE, UK; j.birch2@lse.ac.uk

Received 12 July 2023
Accepted 28 October 2023

To cite: Birch J. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2023-109424

ABSTRACT

Some patients, following brain injury, do not outwardly respond to spoken commands, yet show patterns of brain activity that indicate responsiveness. This is 'cognitive-motor dissociation' (CMD). Recent research has used machine learning to diagnose CMD from electroencephalogram recordings. These techniques have high false discovery rates, raising a serious problem of inductive risk. It is no solution to communicate the false discovery rates directly to the patient's family, because this information may confuse, alarm and mislead. Instead, we need a procedure for generating case-specific probabilistic assessments that can be communicated clearly. This article constructs a possible procedure with three key elements: (1) A shift from categorical 'responding or not' assessments to degrees of evidence; (2) The use of patient-centred priors to convert degrees of evidence to probabilistic assessments; and (3) The use of standardised probability yardsticks to convey those assessments as clearly as possible.

THE SEARCH FOR COGNITIVE-MOTOR DISSOCIATION

Some patients, following brain injury, enter a state of unresponsive wakefulness. Although they have sleep-wake cycles, they give no outward response to any stimulus. This is often known as the 'vegetative state', even though many experts now discourage the use of that term. They discourage it in part because some fraction of patients—and the fraction is unknown—are conscious, experiencing subjects, unable to produce any behavioural report of their experiences. The clinical name for this condition is 'cognitive-motor dissociation' (CMD).[1] Informally, it is often described as 'covert consciousness'.[2]

The risks of failing to diagnose CMD are extremely serious. Some of these risks can be mitigated by low-cost precautions that could be taken with all unresponsive patients, such as administering pain relief and explaining what is happening. Yet it would be a mistake to think accurate diagnosis is therefore unimportant. A diagnosis of CMD is likely to influence life-or-death decisions about the patient's best interests.[3–5]

In the first 2 weeks after a serious brain injury, the patient's surrogate decision makers, in discussion with clinicians, will typically face the terrible decision of whether or not to withdraw life-sustaining treatment.[6] The surrogate decision makers are usually family members, with some exceptions,[7] so I will say 'family' in what follows. Evidence of CMD could have a major influence on their decision, particularly if CMD turns out to be linked to a higher probability of recovery, something that is currently unclear.[3] A Canadian study found withdrawal of treatment to be by far the largest cause of hospital mortality in patients with traumatic brain injury, accounting for 70.2% of deaths.[8] The concern that outwardly unresponsive patients are often written off much too quickly, leading to a 'self-fulfilling prophecy' of no recovery, is a major motivation for research into CMD.[1 9]

Later on, if the patient stabilises in an unresponsive condition, the family—in jurisdictions where this is legal—will face a decision that is yet more terrible: that of whether to withdraw clinically assisted nutrition and hydration, leading eventually to death at a slow speed that is often highly distressing to witness.[10 11] Again, evidence of CMD could play a major role in that decision, though the role it plays will depend on the family's view of the patient's values and interests. For some, the idea of withdrawing nutrition and hydration from a potentially conscious patient is too abhorrent to contemplate. For others, the greater fear is that the patient will continue to live in a way they would experience as a form of torment.

Given the gravity of these decisions, there is a pressing need for reliable ways of diagnosing CMD as early as possible, ideally in the intensive care unit (ICU), in the first few days after admission to hospital. One promising approach, and the focus of a great profusion of recent research, involves the use of electroencephalogram (EEG) recordings of brain activity (a recent review surveyed 119 recent articles on this topic).[12] The guiding thought is that, even when the patient cannot respond behaviourally to stimulation, their patterns of brain activity might still respond in a way that contains clues as to the presence or absence of experience.

There are many techniques in development of this general type, all (it is fair to say) at an early stage. None has yet been rolled out to widespread clinical use. The technology is moving fast, however, and the European Academy of Neurology already recommends the use of EEG and functional magnetic resonance imaging (fMRI) techniques 'whenever feasible' and proposed that patients should be diagnosed as having the 'highest level' of consciousness indicated by behaviour, EEG or fMRI.[13]

Among EEG-based methods, particular excitement surrounds the idea of using machine learning to infer responsiveness to spoken commands from the EEG.[14] This is an important emerging case of the clinical application of artificial intelligence (AI). Yet these machine learning techniques have high false discovery rates, raising a serious problem of inductive risk. In reaching any categorical judgement about whether the patient is responding, the risk of misattributed responsiveness must be

balanced against that of missed responsiveness. This balancing involves value-judgements about the comparative seriousness of the two types of error.[4 9]

I will argue that the value-judgements involved in categorical assessments are not inherently a problem, but they can lead to problems if the values in question are misaligned with the patient's own values. To secure greater sensitivity to the patient's values, what is needed, I argue, is a procedure for generating case-specific probabilistic assessments that can be communicated clearly to the patient's family. I will put forward a possible procedure built around three proposals: (1) A shift from categorical 'responding or not' assessments to degrees of evidence; (2) The use of patient-centred priors to convert degrees of evidence to probabilistic assessments; and (3) The use of standardised probability yardsticks to convey those assessments as clearly as possible to the patient's family.

The problem here is fundamentally a problem of *inductive risk*: a risk associated with moving from probabilistic evidence to the acceptance or rejection of a hypothesis. This inductive step always brings with it the possibility of error, requiring careful evaluation of the comparative seriousness of different types of error. Accordingly, the article aims to build on previous discussions of inductive risk in the management of disorders of consciousness[4 9] which did not zoom in specifically on the complications introduced by machine learning. It is also a contribution to a growing literature on the management of inductive risk in medicine[15–25] and in machine learning.[26 27] The overall message will be that the clinical application of AI to the diagnosis of CMD is a source of new risks and new opportunities. The risk is that contentious value-judgements will be buried too deeply to allow room for input or scrutiny by the patient's family. The corresponding opportunity is that, with the right design, an AI product could enable human clinicians to do a *better* job of evaluating and communicating diagnostic uncertainty than they do currently.

BACKGROUND UNCERTAINTY: THE LINKS BETWEEN RESPONSIVENESS AND CONSCIOUSNESS

When thinking about diagnostic uncertainty, it can help to introduce a distinction between the 'background' and the 'foreground'. Background uncertainty concerns the relevance of a proposed biomarker to the condition we are trying to diagnose (eg, 'Is fever evidence of COVID-19?'). Foreground uncertainty concerns whether or not the biomarker is present or absent in a particular case (eg, 'Is *this* a fever?'). Our focal condition is CMD, and our focal biomarker will be an EEG signature of neural responsiveness to spoken commands.

I want to focus primarily on foreground uncertainty: uncertainty about whether the EEG shows a patient is responding to commands or not. But to put that discussion in context, we should also note three important sources of background uncertainty regarding the relationship between that marker and consciousness.

First, it is far from certain that neural responsiveness to commands, when present, implies conscious experience. There is some evidence that task-relevant responses can be elicited by spoken commands during sleep, when the subject is unconscious according to their own subsequent report.[i28] I think we should grant, however, that neural responsiveness *raises the probability* of conscious experience, since it is more likely to be observed if the patient is conscious than if the patient is unconscious. A strong and implausible epiphenomenalism about conscious experience may deny this, but it can be granted by any view on which, in healthy controls, conscious experience has a causal role in mediating motor responses to linguistic commands.

Second, any inference from the *absence* of responsiveness to the *absence* of consciousness is tendentious. There are many reasons why a conscious patient might fail to respond neurally to spoken commands, including effects of sedation and deficits of attention, memory and linguistic comprehension.[1 3]

Third, even granting that responsiveness raises the probability of consciousness, it leaves many questions open regarding the *form* of the subject's conscious experiences. Current clinical practice involves distinguishing different conscious 'levels': a typical taxonomy includes coma, unresponsive wakefulness, minimally conscious state-minus (MCS−), minimally conscious state-plus (MCS+), confusional state, cognitive dysfunction and full recovery.[1] Finding responsiveness in a patient's EEG does not tell us where to put the patient on these scales (eg, whether to reclassify them as MCS−).

Moreover, CMD casts some doubt on the very idea of a 'levels' framework. The conscious states of unresponsive patients vary a great deal, with some having experiences closely akin to those of a healthy adult, and others having highly degraded, fragmentary, fleeting experiences. Over the long-term, we will surely need a richer framework for thinking about these cases, with many different dimensions of variation, and a shift from 'levels of consciousness' to multidimensional consciousness profiles.[29] Merely finding responsiveness leaves us in the dark as to the patient's consciousness profile. This background uncertainty needs to be included in conversations about prognosis and treatment that follow any evidence of responsiveness. Although my focus in what follows will be on foreground uncertainty, we also need to consider the question of how best to communicate background uncertainty (an issue revisited later in the discussion).

FOREGROUND UNCERTAINTY: IS THE PATIENT RESPONDING AT ALL?

Background uncertainty arises even when we are certain that the putative biomarker is present, but there is also uncertainty about whether the biomarker is there at all. The appearance of responsiveness in the EEG could conceivably be a chance pattern or a statistical artefact.

This possibility has been a source of controversy. In a 2013 study published in the *Lancet*, Cruse *et al* used a machine learning method to analyse EEG data from 16 patients with disorders of consciousness and 12 healthy controls.[30] Across a series of blocks, subjects were instructed to imagine either closing their right hand into a fist or wiggling their toes. The machine learning algorithm, a support vector machine classifier, was tasked with inferring the command given in each block from the EEG response. Significantly above-chance performance by the classifier was interpreted as evidence of responsiveness. The headline finding: 3/16 outwardly unresponsive patients were responding to commands.

In a critique, Goldfine *et al* took issue with the statistical techniques used to detect responsiveness.[31] I will not go into detail here, because to do so would distract from the main case study in the next section. In brief, Goldfine *et al* criticised the method of cross-validation used, the way p-values were calculated, and the chosen significance threshold of $p<0.05$. As has often been

[i]This evidence is itself subject to substantial foreground uncertainty (ie, are the sleeping patients *really* responding?), since Kouider *et al* measured responsiveness using readiness potentials, which have been the targets of methodological criticism.

noted,[32] this threshold leads to high false discovery rates (the false discovery rate is the number of *incorrect* rejections of the null hypothesis divided by the *total* number of rejections). Goldfine *et al* dramatically showed that, when their preferred method of cross-validation was used, when a different method was used to calculate the p-value, and when a correction for multiple comparisons (specifically, a Benjamini-Hochberg correction) was applied to lower the significance threshold, the headline result disappeared: there was no finding of responsiveness in any of the patients.

Cruse *et al* replied combatively, accepting none of the criticisms.[33] My aim is not to referee the dispute here. I will restrict myself to two comments. First, the dispute very clearly shows how the frequency with which responsiveness is detected depends quite sensitively on methodological choices; to acknowledge this is not to take sides on the issue of whose choices were correct. Second, the disagreement suggests a difference in attitude towards the risk of misattributed and missed responsiveness. Goldfine *et al* were concerned by a high false discovery rate and sought ways of controlling it. Cruse *et al* feared that what they describe as 'conservative corrections' would drive up the rate of missed responsiveness. These issues will resurface when we turn to our focal example, a more recent study that takes the criticisms of Goldfine *et al* at least partly on board.

MANAGING INDUCTIVE RISK: HOW VALUES DRIVE METHODOLOGICAL CHOICES

With these issues in mind, I want to examine a high-profile EEG study by Claassen *et al*, published in 2019 in the *New England Journal of Medicine*.[14] This was a ground-breaking study of CMD in an ICU setting, involving an unprecedented sample size of 104 patients with disorders of consciousness. My aim will be to tease out the ways in which value-judgements about the comparative seriousness of missed and misattributed responsiveness shaped methodological decisions.

Before going deeper into methodological details, we should note that the terms 'false positive' and 'false negative' can lead to confusion. The term 'false positive' is sometimes used to describe a single incorrect guess by a classifier, but it is also sometimes used to describe a situation in which a classifier is judged to be performing at above-chance level when the patient is not in fact responding. That is why I favour the term 'misattributed responsiveness' to describe the latter type of situation, and the term 'missed responsiveness' to describe a situation in which a patient is responding but this is not detected in the form of above-chance classifier performance.

In the Claassen *et al* study, a support vector machine classifier was tasked with guessing the spoken commands given to a patient using only an EEG recording of that patient. The commands given were 'keep opening and closing your right[/left] hand' and 'stop opening and closing your right[/left] hand'. The algorithm was trained separately on each patient over the course of 6 blocks of 8 trials each (ie, 48 trials). This is called an 'individualised classifier' approach, since the classifier is trained anew on every patient's personal EEG data. This strategy can be contrasted with a 'general classifier' approach that seeks to generalise from a training set of patients to a new patient.

For each patient, the classifier's performance was evaluated by comparing its guesses about the spoken commands (inferred from the EEG) to the actual commands.[ii] The headline result: in 16/104 patients, significantly above-chance classifier performance was obtained, leading the authors to the striking conclusion that 'of the 104 patients, 16 (15%) had cognitive–motor dissociation detected on at least one recording'.[14]

For any study of this type, researchers face many difficult methodological choices. I will focus on four:

A. What significance threshold will be used to assess 'better than chance' performance by the classifier?
B. How many EEG recordings will be made of each patient and at what points in the timeline (eg, in the acute, subacute or chronic phase)?
C. How will adjustments to the significance threshold be made for multiple recordings of a single patient, and how will the point in the timeline affect the statistical approach taken (eg, will we be less conservative for patients in the acute phase)?
D. How will adjustments to the significance threshold be made to control the false discovery rate in the whole sample of patients?

All these choices have implications for the likely rates of misattributed and missed responsiveness. Let us consider how Claassen *et al* handled them.

Regarding (A): Claassen *et al* used a standard *p*-value threshold of 0.05 to assess whether the classifier was performing significantly better than chance. A value of $p=0.05$ implies a 0.05 probability of the observed level of classifier performance being achieved by chance, without real responsiveness. As noted earlier, a threshold of $p<0.05$ is well known to create a risk of a high false discovery rate when many tests are conducted. That needs to be kept in mind as we consider (B)–(D). How did Claassen *et al* try to manage that risk?

Regarding (B) and (C): in keeping with the aim of developing a diagnostic tool for use in the first few days after brain injury, patients were initially tested between 3 days and 10 days after injury. For some but not all patients, Claassen *et al* took multiple recordings (with an overall median of two recordings per patient). If this type of procedure becomes a widely used diagnostic tool, clinicians will often want to take multiple recordings, because a single recording carries a high risk of missed responsiveness. A key advantage of bedside EEG in the ICU over fMRI outside the ICU is that the former (in addition to being safer and faster) allows for repeated measurement.[1] Yet doing multiple recordings drives up the chance of misattributed responsiveness.

There are two well-known statistical techniques for managing this risk: the Bonferroni procedure and the above-mentioned Benjamini-Hochberg procedure, recommended by Goldfine *et al*.[31] The two procedures control different quantities. The Bonferroni procedure controls the overall rate of false positives (type I errors) in a series of tests. If we apply this procedure, we can be reassured that the overall chance of there being a misattribution of responsiveness to a given patient is below the desired threshold (eg, 0.05). Yet reassurance about the chance of a false positive is bought at the cost of driving up the rate of false negatives. The Benjamini-Hochberg procedure, meanwhile, controls the false discovery rate (the fraction of cases of misattributed responsiveness among the total number of positive tests). After

[ii]The classifier's performance was evaluated by the area under the receiver operating curve (AUC). For patients who are not responding, the classifier is expected to perform at chance level, which corresponds to an AUC neither significantly greater than nor significantly less than 0.5. For patients who are responding, the classifier is expected to perform at a significantly above-chance level, corresponding to an AUC significantly greater than 0.5. In the main text I will simply refer to the 'performance level' of the classifier.

applying this procedure, we can regard each 'positive' recording, indicating apparent responsiveness, as having at most a 5% chance of being a misattribution.

Claassen *et al* chose the Benjamini-Hochberg procedure. This is the less stringent of the two options, since controlling the false positive rate is more stringent than controlling the false discovery rate when there are many comparisons. The trouble here is that, in a clinical setting, it is surely the overall chance of a misattribution *for this patient* (ie, the variable controlled by the Bonferroni procedure) that we most want to control. If we simply control the false discovery rate among EEG recordings, we are still faced with a situation where the chance of a misattribution happening at some point, for any given patient, becomes very high as the number of EEG recordings conducted on that patient goes up. To give an extreme illustration: if we make 100 recordings of the same patient, a misattribution somewhere in the sequence is very likely, even if we apply the Benjamini-Hochberg procedure to hold the false discovery rate at 5% or lower.

Regarding (D): Beyond the correction just noted for individual patients who were recorded multiple times, Claassen *et al* did not make any further downward adjustments of the significance threshold to control the overall rate of false discoveries in the population as a whole. So, the overall false discovery rate was likely to be high.

Strikingly, Claassen *et al* write in their supplementary information that, since 104 patients were studied, 'it is likely that amongst the 16 CMD patients, five were classified as CMD because of statistical fluctuations rather than actual spoken command following' ([14], supplementary information, page 13). In fact, inferences from the *p*-value to the false discovery rate are more complicated than this quotation makes it sound. What the authors should have said is that, in a population of 104 patients with no cases of true responsiveness, we would expect about five cases of misattributed responsiveness. However, it does not follow that in the real study population, where some cases of true responsiveness very probably existed, five were likely to have been misattributions. Further assumptions about the base rate of the target property (ie, responsiveness) are needed to estimate the false discovery rate. That said, the authors are clearly right to flag the possibility of a high false discovery rate, given their methodological choices.

Could Claassen *et al* have controlled the overall false discovery rate by pushing the *p*-value threshold below 0.05? By way of analogy, this is an orthodox approach in genome-wide association studies (GWASs), which also involve many separate tests for statistical relationships (but across many genes rather than many patients). Since the 2000s, the norm in this area has been to use a *p*-value threshold of $p < 5 \times 10^{-8}$.[34 35] By contrast, Claassen *et al* applied the Benjamini-Hochberg procedure to multiple recordings *from the same patient*, but not to the *whole set* of tests across *all* patients. The analogy in a GWAS would be to control the false discovery rate for repeated tests of the same gene, without controlling the overall false discovery rate across the whole genome. However, it is not clear that controlling the false discovery rate across the whole population would have been the right approach here, for it would have made the threshold for a diagnosis of responsiveness sensitive to the population size. As already emphasised in the discussion of (B) and (C), surely what we really want is a diagnosis for *each individual patient* that has a low chance of error, irrespective of how many other patients there are or how many times each patient is recorded.

In sum, Claassen *et al* chose to set a reasonably easy-to-clear bar for statistical significance ($p < 0.05$), chose to adjust it downwards for multiple comparisons only in a limited and partial way, and chose to accept—and openly acknowledge, although in supplementary information—a high false discovery rate.

These choices reflect implicit value-judgements by the researchers. That is not intended as a criticism, because I regard these value-judgements as both unavoidable and potentially benign. The ICU is not a genomics lab, and the consequences of error are very different in the two contexts. Claassen *et al* were quite clearly, and understandably, worried about the risk of missed responsiveness. As Fins and Bernat have emphasised, the 'ethical importance of avoiding type II error: failing to identify consciousness when it is present' looms large for researchers in this area.[2] This concern drives methodological choices that prioritise avoiding cases of missed responsiveness, while expressing a more permissive attitude towards cases of misattributed responsiveness.

A PROBLEM: NEGLECTING THE PATIENT'S OWN VALUES

Here is the story so far: multiple EEG testing of each patient, plus a relatively high threshold for significance ($p < 0.05$), is a recipe for a high rate of misattribution. This is exemplified by the Claassen *et al* study, in which the authors themselves estimated a false discovery rate of 5 in 16. While this may be an underestimate or an overestimate, the point of relevance here is that the authors, having reached that estimate, were willing to accept it. There are frequentist strategies for controlling the false discovery rate, but researchers in this area are *understandably* hesitant to use them (except in limited, partial ways) because they are 'conservative' and allow the rate of missed responsiveness to rise in an uncontrolled way—and because the normative, clinical importance of the research speaks strongly against a disregard for the risk of missed responsiveness. A special concern for the risk of missed responsiveness is widespread among researchers working in this area and shapes their methodological choices.

Is this a problem? Clinical research and practice cannot avoid value-judgements altogether. To worry about missed responsiveness more than misattributed responsiveness is a value-judgement, but we may well be tempted to regard it as a benign, well-founded one.

In my view, however, a problem remains, even if we agree with all the relevant value-judgements. The real danger is not that of value-judgements being made in scientific research (this is normal and unavoidable) but that these value-judgements will be made in an insufficiently inclusive and context-sensitive way. The value-judgements are being made in a one-size-fits-all manner by researchers, or software designers, without involvement of families. Families simply receive a result—*the patient is responding*/*not responding*—without an opportunity for input into the underlying value-judgements that shaped the methodological choices leading to this result.

This is a problem even now, since results obtained in research studies are sometimes shared with the patient's family and so already inform decision-making in some cases.[3 36] But it has the potential to become a much larger problem if the approach is rolled out to widespread clinical use, as the European Academy of Neurology has urged. Value-judgements about the comparative seriousness of missed and misattributed responsiveness should, as far as possible, properly involve the patient's surrogate decision makers.

One might object: would *anyone* really disagree that missed responsiveness is much worse than misattributed responsiveness? But it is not that simple. There is a wide range of reasonable reactions to evidence of covert awareness in a patient. The actual pattern of variation in family reactions is still poorly understood,

calling for further research.[37–41] In broad terms, though, we can expect that some families will hope for evidence of responsiveness, regarding it as a sign of incipient recovery, whereas others will fear it, seeing it as a sign that the patient may be suffering and experiencing a low quality of life.[42] For this latter group, a misattribution of responsiveness could conceivably tip the balance in favour of withdrawing treatment.

Meanwhile, in cases where the family hopes for evidence of continuing awareness, a misattribution of responsiveness may lead to false hope: hope that a given level of recovery is in fact realistic, when other evidence suggests it is not.[4] That false hope may be a curse rather than a blessing. Given the current legal framework surrounding end-of-life decisions in most jurisdictions, families are often put in an agonisingly difficult position. There may be only a narrow window in which life-sustaining treatment such as mechanical ventilation can be withdrawn (if this is in the patient's best interests) before the patient's condition stabilises, resulting in a situation where a quick death is no longer legally possible (the path from withdrawal of clinically assisted nutrition and hydration to death is distressingly long by comparison). Misattributed responsiveness could conceivably lead to that window being missed—it could lead to patients being kept alive over the long term when their prospects of recovery to a level they would themselves value are bleak. Kitzinger and Kitzinger have urged clinicians to take this risk seriously.[6] Values in this area vary a great deal. People have very different concepts of what counts as a good recovery, and some patients would want to be kept alive even if that level of recovery was very unlikely, whereas others would not.[3]

In the US context, there is also a further legal complication: courts are far less likely to grant approval for treatment withdrawal in cases where a patient is diagnosed as minimally conscious rather than 'vegetative', essentially forcing patients to be kept alive regardless of their values or wishes.[9] This adds an extra cost to misattributions of responsiveness. In the UK, where court approval is not always needed to withdraw clinically assisted nutrition and hydration, clinical guidance avoids placing such enormous weight on the boundary between minimally conscious and 'vegetative', correctly recognising this to be subject to great uncertainty.[43]

In the absence of a clear advance directive by the patient themselves, the patient's family is generally recognised to be the best (though fallible) way to access the patient's own values. And yet, one can hardly go to the patient's family and ask: 'Should I apply a Bonferroni correction to your relative's EEG data, or a less conservative correction?' *Error rates are controlled by deeply buried methodological details that a patient's relatives will typically be unable to understand*. Yet they will keenly feel the consequences of those methodological choices, because they will be heavily involved in making life-or-death decisions about the patient in which those choices may tip the balance of considerations, either for the family or for a court.

Given this, should clinicians communicate their best estimates of the false discovery rate to the patient's family? Should they say: 'We expect there is a high (but difficult to estimate precisely) false discovery rate associated with this procedure'? Edlow and Fins seem to have something like this in mind when they write 'the possibility of a false-positive result must be considered by clinicians and clearly communicated to families' and, later, that 'it is ethically appropriate to share single-subject data if families are fully informed of the performance characteristics of the assessments, such as their sensitivity and specificity'.[3] Yet this would not adequately solve the problem, and might even make things worse, because false discovery rates are easily misinterpreted. They have significant potential to confuse, alarm and mislead. Not all particular cases are equal, and an expected false discovery *rate* of *x*% does not equate to a *x*% *chance* of a particular case being a false discovery. The EEG may provide extremely strong evidence of responsiveness in some specific cases and extremely weak evidence in other cases. But this further complication—that is, the strength of evidence varies from one case to the next—points us in the direction of a possible solution.

COMMUNICATING UNCERTAINTY, PROPOSAL 1: DEGREES OF EVIDENCE

There may be a temptation to react to a high false discovery rate simply by lowering the required p-value, mirroring standard practice in GWASs, and also mirroring calls in psychology for a lowering of the standard p-value threshold to $p<0.005$.[32] Yet we would still be using a single, one-size-fits-all threshold to assess whether the patient is 'responding' or not. Ultimately, this is misleading. The reality is that classifier performance delivers evidence of responsiveness of continuously varying strength. If the strength of evidence can be conveyed to clinicians and the patient's family, it will enable better-informed decisions.

But how to do this? Our problem arises in part from the use of frequentist methods to interpret EEG recordings. Might the problem be at least partly addressed by incorporating Bayesian ideas? Bayesian approaches are no panacea for deep problems of inductive risk and uncertainty communication, but I believe they can help. There are, in particular, two Bayesian ideas that can help: (1) EEG data do not directly support a yes/no verdict on questions of responsiveness, but rather provide a quantitative degree of evidence; (2) Converting degrees of evidence into probabilistic assessments requires consideration of prior probabilities. The following two proposals should be understood as part of a single package that rests on these ideas. By contrast, Proposal 3 has the status of a standalone extra that could be pursued in combination with any strategy for estimating probabilities.

The first proposal is to shift from p-values to Bayes factor bounds. The lowest p-value threshold at which the algorithm's performance becomes 'significantly' above-chance is a continuous variable that provides some insight into the strength of evidence of responsiveness in the present patient. However, the relationship between p-values and strength of evidence is not straightforward. The p-value imposes an *upper bound* on the Bayes factor, a formal measure of the strength of evidence. In other words, given a certain p-value, there is an upper limit on how much evidence against the null hypothesis a data set can provide. This Bayes factor upper bound (BFB) is typically given by:

$$\mathrm{BFB} = \frac{1}{-ep \log p}$$

where p is the p-value.[44] In the case of multiple recordings of the same patient, Bayes factors can be multiplied to give an upper bound on the total evidence against the null hypothesis provided by the series of tests.

Benjamin and Berger recommend the increased use of BFBs in science as a response to the replication crisis.[44] My first proposal is that BFBs also have an important role in the design of medical AI. In cases where a diagnostic package delivers a p-value as its primary output, the p-value should be converted to a BFB.

An obvious problem arises: doctors and patients' families will not typically find the BFB any more intuitive to interpret than the p-value. So, we cannot stop here. What we need is to

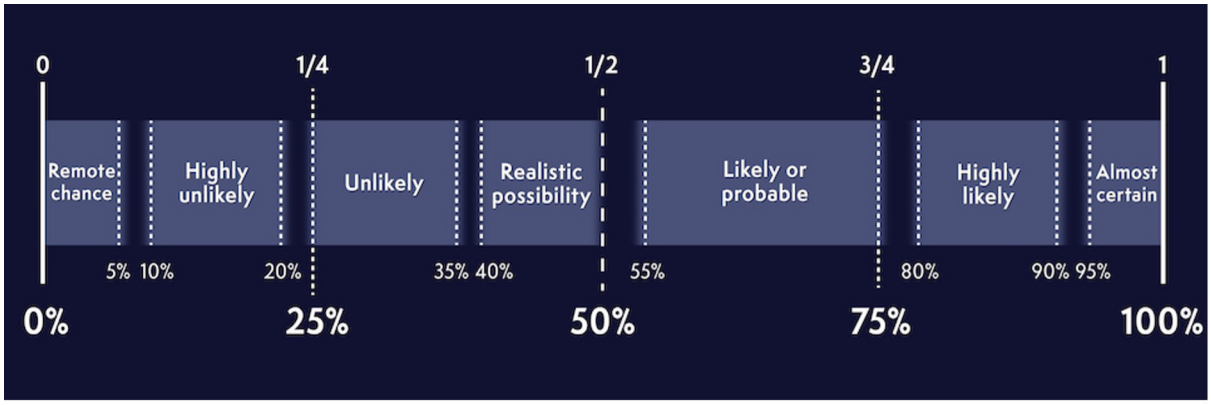

Figure 1 The Professional Head of Intelligence Assessment (PHIA) probability yardstick. This figure can be found widely in UK government documents (public sector information licensed under the Open Government Licence v3.0).

transform the BFB into a more useful output: an assessment of the probability that the patient is responding. To do this, we need to incorporate prior probabilities.

COMMUNICATING UNCERTAINTY, PROPOSAL 2: PATIENT-CENTRED PRIORS

To move from a BFB to an upper bound on the odds that the patient is responding to commands (the 'odds upper bound', or OB), we need prior probabilities:

$$\text{OB}\left(\textit{Responding}\right) = \text{BFB} \times \frac{\text{Prior}\left(\textit{Responding}\right)}{\text{Prior}\left(\textit{Not responding}\right)}$$

But how to decide these priors? One option would be to set our priors in line with base rates. However, at present, our uncertainty about individual cases of CMD percolates up to uncertainty about the base rate. For example, we could use the Claassen *et al* data to estimate a base rate of about 10% among unresponsive wakeful patients (ie, 16/104, minus about five probable misattributions), but that estimate would itself be subject to substantial uncertainty. We can hope that, over the long run, we can be more confident (eg, if many different studies, with similarly large samples but substantially different methodologies, converge on a similar base rate).

Remember, though, that we are thinking here about how to use EEG data to inform decisions about a *specific patient*, and the clinicians treating them will have lots of other background information that might relevantly shape the prior probability of responsiveness in this patient, beyond just base rates. They will have, for example, information about whether the patient is responding behaviourally, and about the extent of their brain injury. The priors for a specific patient call for a holistic expert judgement, a judgement that considers not just base rates but the wider context.

So, a natural suggestion is that a good diagnostic algorithm will include an opportunity for clinicians to enter their all-things-considered priors that the patient will respond to commands, given everything else that is known about them. The algorithm will then combine those priors with the BFB to generate an OB. A probability is somewhat easier to interpret than an odds bound, so it will be helpful if the formula also converts the odds to a probability by applying the formula:

$$\text{Probability Upper Bound}\left(\textit{Responding}\right) = \frac{\text{OB}\left(\textit{Responding}\right)}{1+\text{OB}\left(\textit{Responding}\right)}$$

This proposal comes with an associated risk: a probabilistic version of the 'self-fulfilling prophecy', whereby a clinician gives a patient an extremely low prior probability of responding. If the prior is set low enough, even very strong evidence of responsiveness will not be enough to render it more likely than not that the patient is responding. For example, suppose, due to the severity of the patient's injury, a clinician inputs a prior probability of responding below 0.001. This might be described as a 'dogmatic' prior: a prior so low that even very strong evidence of responsiveness may fail to move the probability upper bound to 0.5 or higher. We need to guard against this risk of dogmatic priors. After all, the development of this technology is driven in the first place by a fear that clinicians are too willing to 'write off' patients as unresponsive when they are not. My proposal provides a mechanism through which this very fear could be realised. It may at first sight seem rather idealistic to think doctors could be entrusted with the task of inputting reasonable priors.

This brings us to the nub of the problem: in setting patient-centred priors, how are we to integrate the expert judgement of clinicians, which inevitably and appropriately relies on low-tech behavioural evidence, with a need for openness to the possibility of covert consciousness that clinicians are unable to detect without technological assistance? I hope to start a debate on this issue but do not claim to have a perfect solution. I propose there should be clinical guidelines requiring clinicians input priors that fall within an *appropriately open-minded* range, after discussion with all parties involved in the clinical care of the patient, including the patient's family. These priors should be non-dogmatic and anchored on recommendations from professional bodies based on the overall evidential picture concerning base rates.

It would be premature for me to specify exact thresholds for what counts as a 'dogmatic' prior (the example of 0.001 is not a proposed threshold; it is just an illustrative example of a very low prior). What is needed is clear guidance from professional bodies further down the road, and it is not my place to prejudge that guidance. But as a tentative proposal with the aim of provoking debate, I suggest that the *highest realistic estimates* of the base rates of covert awareness, according to current evidence, should be used as an anchoring point, and that departures from that anchoring point should be small and justified by clinical evidence specific to the current patient. *Prior probabilities that fall far below existing published estimates of the base rate (as in the imagined case of a clinician who inputs 0.001) should be avoided*. In assessing the 'highest realistic estimate', professional bodies should consider all relevant evidence from consciousness science, casting a wide net. Crucially, theories of consciousness that imply that conscious experience may persist

despite very extensive cortical damage, such as the midbrain-centric theories of Bjorn Merker and Jaak Panksepp, should be given careful consideration in this process.[45–49]

This proposal does not remove the need for value-judgements. Indeed, my proposal that clinicians should anchor their priors to the *highest* realistic estimate of the base rate implicitly involves a value-judgement. I am proposing, in effect, that it would be worse to underestimate the base rate than to overestimate it, and some may disagree. However, I do think the problem of neglecting the patient's own values can be mitigated, provided the patient's family is involved in the discussions through which the clinician decides on an appropriate prior.

COMMUNICATING UNCERTAINTY, PROPOSAL 3: PROBABILITY YARDSTICKS

Imagine the following scenario: a clinician inputs patient-centred priors into a software package, the package calculates an upper bound on the strength of evidence of responsiveness from EEG data, and reports back an upper bound on the probability that the patient is responding. This report may still be very difficult for the patient's family to interpret, so our framework is not yet complete. How can this upper bound be communicated sensitively to the patient's family, so as to put them in a better position to make the decisions that lie before them?

Without standardised language, numerical probabilities can be converted to ordinary language terms in many different ways.[50] A 'probabilistic yardstick' aims to solve this problem by providing a standardised protocol for assigning verbal, qualitative labels to probability ranges. An influential example is the Professional Head of Intelligence Assessment probability yardstick, widely used in UK government circles (figure 1). This yardstick maps the terms 'remote chance', 'highly unlikely', 'unlikely', 'realistic possibility', 'likely/probably', 'highly likely' and 'almost certain' to ranges of probabilities.

I see this as a starting point, but far from a perfect proposal. To require 40% probability before being willing to describe an outcome as a 'realistic possibility' is unwarranted. I would favour the label 'about as likely as not' for the range 45%–55%. Moreover, the word 'likely' covers too big a range, including outcomes that are slightly more likely than not (~55%), outcomes that are moderately likely (~60–70%), and outcomes that have a~75% probability of occurring. Yet this starting point illustrates the general idea. My third proposal is that *standardised yardsticks should be developed, in consultation with clinicians, patients (where possible) and patients' families, for use in cases where diagnostic AI yields quantitative probabilistic outputs*.

Here is a design choice point the proposal raises: who should implement the conversion of quantitative probabilities to qualitative categories? At present, patients are generally sceptical of the idea of AI *replacing* a human clinician in making critical judgements and decisions, such as whether to recall a patient for a biopsy following cancer screening.[27 51 52] To be clear, no part of my proposal involves the AI *deciding* anything. However, I suspect that hiding the raw probability may still raise a problem of trust for some families. There is also some evidence that people with a good level of numeracy *prefer* information about risk to be conveyed to them numerically.[50] A solution would be for the algorithm to output *both* a precise probability upper bound and a suggested qualitative interpretation. For example, the output might read:

> It is at most moderately likely that the patient responded to simple commands during the series of tests performed at [times, dates]. (Estimated probability upper bound: 62%). Please note that this is assessment of the probability of responsiveness, not consciousness. A conscious patient may still fail to respond for many reasons.

Informing the user that we are assessing the probability of *responsiveness*, not consciousness, is crucial. More generally, it is crucial that we do not forget background uncertainty when attempting to communicate foreground uncertainty. A good qualitative interpretation will highlight both kinds of uncertainty. A clinician can then communicate this result in a way appropriate to the patient's family, giving probabilities if they have a good grasp of the concept of probability, and using coarse-grained, qualitative categories if they do not. Patients' families will then be well placed to take this information into account in a way that respects what is known about the patient's values and wishes.

OPPORTUNITIES AND RISKS OF MEDICAL AI

In sum, the use of EEG recordings to detect covert consciousness raises a serious problem of inductive risk, calling for a value-judgement about the comparative seriousness of misattributed and missed responsiveness. Current methods bury value-judgements in 'under-the-hood' methodological choices, opaque to the patient's family. To address this, we should look for ways of incorporating the patient's values transparently (to the extent that they are known by the family) into the management of risk.

My proposal for one way to do this involves three ingredients: (1) A shift from 'responding or not' to *degrees of evidence* quantified by Bayes factor bounds; (2) The use of *patient-centred priors* to convert degrees of evidence to probabilistic assessments; and (3) The use of standardised *probability yardsticks* to convey those assessments clearly to the patient's family, who are best placed to know what the patient would want.

What are the wider lessons of the case for the clinical use of AI? The case highlights a type of risk that is likely to recur in many clinical contexts: the risk of an algorithm encoding implicit value-judgements, such as judgements about the comparative seriousness of false positives and false negatives, that differ from those the patient would want to be made. The same risk arises in the case of cancer screening.[27] It will arise whenever an algorithm is tasked with moving from a native output that is fundamentally probabilistic to a yes/no judgement.

Avoiding AI altogether would not remove that risk, since human clinicians can also make implicit value-judgements in conflict with the patient's own values. That said, patients and their families often trust their clinicians to have their best interests at heart, whereas the same level of trust in AI does not exist.[27 51 52] So, there is a risk of eroding patient trust in cases where AI products are found to be encoding contentious value-judgements.

With this risk comes a corresponding opportunity. Human clinicians often struggle to estimate and communicate uncertainty. This can lead to an exaggerated sense of certainty surrounding early diagnoses in an area where misdiagnosis is easy and common.[9] Poorly designed medical AI could accentuate the problem, if it gives yes/no verdicts on matters as shrouded in uncertainty as the presence or absence of responsiveness to commands. By contrast, well-designed medical AI could help to foster the humility and open-mindedness that is needed in these cases. It can do so by delivering inputs to decision-making that are explicitly probabilistic, while also accompanied by clearly defined qualitative language that the patient's family can more easily understand.

Twitter Jonathan Birch @birchlse

Acknowledgements I am very grateful to Tim Bayne, Liam Kofi Bright, Heather Browning, Katariina Hynninen, Syd Johnson, Anya Plutynski, eva read, Mona-Marie Wandrey and three anonymous reviewers for their comments and advice. I thank Abhinav Jha and Katie Creel for helpful conversations about probabilistic classifiers.

Contributors JB conducted all elements of the research, including writing the article, and is the guarantor.

Funding This research is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant Number 851145.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing is not applicable as no data sets were generated and/or analysed for this study.

ORCID iD
Jonathan Birch http://orcid.org/0000-0001-7517-4759

REFERENCES

1 Edlow BL, Claassen J, Schiff ND, *et al*. Recovery from disorders of consciousness: mechanisms, prognosis and emerging therapies. *Nat Rev Neurol* 2021;17:135–56.
2 Fins JJ, Bernat JL. Ethical, palliative, and policy considerations in disorders of consciousness. *Neurology* 2018;91:471–5.
3 Edlow BL, Fins JJ. Assessment of covert consciousness in the intensive care unit: clinical and ethical considerations. *J Head Trauma Rehabil* 2018;33:424–34.
4 Peterson A, Cruse D, Naci L, *et al*. Risk, diagnostic error, and the clinical science of consciousness. *Neuroimage Clin* 2015;7:588–97.
5 Peterson A, Owen AM, Karlawish J. Translating the discovery of covert consciousness into clinical practice. *JAMA Neurol* 2020;77:541–2.
6 Kitzinger J, Kitzinger C. The 'window of opportunity' for death after severe brain injury: family experiences. *Sociol Health Illn* 2013;35:1095–112.
7 Fins JJ. Disorders of consciousness and disordered care: families, caregivers, and narratives of necessity. *Arch Phys Med Rehabil* 2013;94:1934–9.
8 Turgeon AF, Lauzier F, Simard J-F, *et al*. Mortality associated with withdrawal of life-sustaining therapy for patients with severe traumatic brain injury: a Canadian multicentre cohort study. *CMAJ* 2011;183:1581–8.
9 Johnson LSM. *The Ethics of Uncertainty: Entangled Ethical and Epistemic Risks in Disorders of Consciousness*. New York: Oxford University Press, 2021.
10 Kitzinger C, Kitzinger J. Withdrawing artificial nutrition and hydration from minimally conscious and vegetative patients: family perspectives. *J Med Ethics* 2015;41:157–60.
11 Kitzinger J, Kitzinger C. Deaths after feeding-tube withdrawal from patients in vegetative and minimally conscious States: a qualitative study of family experience. *Palliat Med* 2018;32:1180–8.
12 Bai Y, Lin Y, Ziemann U. Managing disorders of consciousness: the role of electroencephalography. *J Neurol* 2021;268:4033–65.
13 Kondziella D, Bender A, Diserens K, *et al*. European Academy of neurology guideline on the diagnosis of coma and other disorders of consciousness. *Eur J Neurol* 2020;27:741–56.
14 Claassen J, Doyle K, Matory A, *et al*. Detection of brain activation in unresponsive patients with acute brain injury. *N Engl J Med* 2019;380:2497–505.
15 Bavli I, Steel D. Inductive risk and Oxycontin: the ethics of evidence and post-market surveillance of pharmaceuticals in Canada. *Public Health Ethics* 2020;13:300–13.
16 Biddle JB. Inductive risk, epistemic risk, and overdiagnosis of disease. *Perspect Sci* 2016;24:192–205.
17 Bluhm R. Inductive risk and the role of values in clinical trials. In: Elliot KC, Richards T, eds. *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, 2017: 193–214.
18 Douglas HE. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press, 2009.
19 Kostko A. Inductive risks and psychiatric classification. In: Tekin S, Bluhm R, eds. *The Bloomsbury Companion to Philosophy of Psychiatry*. London: Bloomsbury, 2019: 197–216.
20 Kukla R. Infertility, Epistemic risk, and disease definitions. *Synthese* 2019;196:4409–28.
21 Lewens T. The division of advisory labour: the case of 'mitochondrial donation'. *Euro Jnl Phil Sci* 2019;9:1–24.
22 Plutynski A. Safe, or sorry? Cancer screening and Inductive risk. In: Elliott KC, Richards T, eds. *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, 2017: 149–69.
23 Scarantino A. Inductive risk and justice in kidney allocation. *Bioethics* 2010;24:421–30.
24 Stanev R. Inductive risk and values in composite outcome measures. In: Elliott KC, Richards T, eds. *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, 2017: 171–92.
25 Stegenga J. Drug regulation and the inductive risk Calculus. In: Elliott KC, Richards T, eds. *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, 2017: 17–36.
26 Karaca K. Values and Inductive risk in machine learning modelling: the case of binary classification models. *Euro Jnl Phil Sci* 2021;11:1–27.
27 Birch J, Creel KA, Jha AK, *et al*. Clinical decisions using AI must consider patient values. *Nat Med* 2022;28:229–32.
28 Kouider S, Andrillon T, Barbosa LS, *et al*. Inducing task-relevant responses to speech in the sleeping brain. *Curr Biol* 2014;24:2208–14.
29 Bayne T, Hohwy J, Owen AM. Are there levels of consciousness. *Trends Cogn Sci* 2016;20:405–13.
30 Cruse D, Chennu S, Chatelle C, *et al*. Bedside detection of awareness in the vegetative state: a cohort study. *Lancet* 2011;378:2088–94.
31 Goldfine AM, Bardin JC, Noirhomme Q, *et al*. Reanalysis of 'bedside detection of awareness in the vegetative state: a cohort study.' *Lancet* 2013;381:289–91.
32 Benjamin DJ, Berger JO, Johannesson M, *et al*. Redefine statistical significance. *Nat Hum Behav* 2018;2:6–10.
33 Cruse D, Chennu S, Chatelle C, *et al*. Reanalysis of 'bedside detection of awareness in the vegetative state: a cohort study' - authors' reply. *Lancet* 2013;381:291–2.
34 Chen Z, Boehnke M, Wen X, *et al*. Revisiting the genome-wide significance threshold for common variant GWAS. *G3 (Bethesda)* 2021;11:jkaa056.
35 Fadista J, Manning AK, Florez JC, *et al*. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet* 2016;24:1202–5.
36 Fins JJ. *Rights Come to Mind: Brain Injury, Ethics, and the Struggle for Consciousness*. Cambridge: Cambridge University Press, 2015.
37 Andersen LM, Boelsbjerg HB, Høybye MT. Disorders of consciousness: an embedded ethnographic approach to uncovering the specific influence of functional neurodiagnostics of consciousness in surrogate decision making. *Neuroethics* 2021;14:351–6.
38 Boegle K, Bassi M, Comanducci A, *et al*. Informal caregivers of patients with disorders of consciousness: a qualitative study of communication experiences and information needs with physicians. *Neuroethics* 2022;15:24.
39 Kuehlmeyer K, Bender A, Jox RJ, *et al*. Next of kin's reactions to results of functional neurodiagnostics of disorders of consciousness: a question of information delivery or of differing epistemic beliefs *Neuroethics* 2021;14:357–63.
40 Peterson A. How will families react to evidence of covert consciousness in brain-injured patients. *Neuroethics* 2021;14:347–50.
41 Schembs L, Ruhfass M, Racine E, *et al*. How does functional neurodiagnostics inform surrogate decision-making for patients with disorders of consciousness? A qualitative interview study with patients' next of kin. *Neuroethics* 2021;14:327–46.
42 Wilkinson DJ, Kahane G, Horne M, *et al*. Functional neuroimaging and withdrawal of life-sustaining treatment from vegetative patients. *J Med Ethics* 2009;35:508–11.
43 Royal College of Physicians and British Medical Association. Clinically-assisted nutrition and hydration (CANH) and adults who lack the capacity to consent: guidance for decision-making in England and Wales. 2018. Available: https://www.bma.org.uk/canh
44 Benjamin DJ, Berger JO. Three recommendations for improving the use of P-values. *Am Stat* 2019;73:186–91.
45 Merker B. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav Brain Sci* 2007;30:63–81.
46 Panksepp J. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press, 1998.
47 Panksepp J. Affective consciousness: core emotional feelings in animals and humans. *Conscious Cogn* 2005;14:30–80.
48 Panksepp J. The basic emotional circuits of mammalian brains: do animals have affective lives? *Neurosci Biobehav Rev* 2011;35:1791–804.
49 Panksepp J, Fuchs T, Garcia VA, *et al*. Does any aspect of mind survive brain damage that typically leads to a persistent vegetative state? Ethical considerations. *Philos Ethics Humanit Med* 2007;2:32.
50 Lipkus IM. Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. *Med Decis Making* 2007;27:696–713.
51 Ongena YP, Yakar D, Haan M, *et al*. Artificial intelligence in screening mammography: a population survey of women's preferences. *J Am Coll Radiol* 2021;18:79–86.
52 Temple S, Rowbottom C, Simpson J. Patient views on the implementation of artificial intelligence in radiotherapy. *Radiography (Lond)* 2023;29 Suppl 1:S112–6.