Draft paper, 12 July 2023

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,
London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, UK.
j.birch2@lse.ac.uk
http://personal.lse.ac.uk/birchj1

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 (EEG) 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.

Key words: disorders of consciousness, cognitive-motor dissociation, inductive risk, AI and medicine.

1. 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. This clinical name for this condition is "cognitive-motor dissociation" (CMD) (Edlow et al. 2021). Informally, it is often described as "covert consciousness" (Fins and Bernat 2018).

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 (as urged by Fins and Bernat 2018) 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 (Edlow and Fins 2018; Peterson et al. 2015, 2020).

In the first two 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 (Kitzinger and Kitzinger 2013). The surrogate decision makers are usually family members, with some exceptions (Fins 2013), 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 (Edlow and Fins 2018). A Canadian study by Turgeon et al. (2011) 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. 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 (Johnson 2022; Edlow et al. 2021).

Later on, if the patient stabilizes 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 for the patient's family (Kitzinger & Kitzinger 2015, 2018). 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 (119 recent articles on this are reviewed in Bai et al. 2021). 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 fMRI techniques "whenever feasible" and proposed that patients should be diagnosed as having the "highest level" of consciousness indicated by behaviour, EEG or fMRI (Kondziella et al. 2020; Edlow et al. 2021).

Among EEG-based methods, particular excitement surrounds the idea of using machine learning to infer responsiveness to spoken commands from the EEG (Claassen et al. 2019). This is an important emerging case of the clinical application of 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 (Peterson et al. 2016; Johnson 2022).

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 construct 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 standardized probability yardsticks to convey those assessments as clearly as possible to the patient's family.

The article aims to build on previous discussions of inductive risk in the management of disorders of consciousness (especially Peterson et al. 2015; Johnson 2022), 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 more generally (Bavli & Steel 2020; Biddle 2016; Bluhm 2017; Douglas 2009; Kostko 2019; Kukla 2019; Lewens 2019; Plutynski 2017; Scarantino 2010; Stanev 2017; Stegenga 2017) and in machine learning (Karaca 2021; Birch et al. 2022). 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 highly 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.

2. 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 (e.g. "Is fever evidence of COVID-19?"). Foreground uncertainty concerns whether or not the biomarker is present or absent in a particular case (e.g. "Is *this* a fever?"). Our focal condition is CMD, and our focal biomarker will be neural responsiveness to spoken commands.

I want to focus primarily on foreground uncertainty: uncertainty about whether 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 (Kouider et al. 2014).[1] 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 (Edlow and Fins 2018; Edlow et al. 2021).

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 (UWS), minimally conscious state-minus (MCS-), minimally conscious state-plus (MCS+), confusional state, cognitive dysfunction, and full recovery (Edlow et al. 2021). Finding responsiveness in a patient's EEG does not tell us where to put the patient on these scales (e.g. 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

[1] This evidence is subject to substantial foreground uncertainty (i.e. are the sleeping patients *really* responding?), since Kouider et al. measured responsiveness using readiness potentials, which have been the targets of methodological criticism (see Section 3).

these cases, with many different dimensions of variation, and a shift from "levels of consciousness" to multidimensional consciousness profiles (Bayne et al. 2016). Merely finding responsiveness leaves us in the dark as to the patient's consciousness profile.

3. 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. Cruse et al. (2011), in a study published in the *Lancet*, used a machine learning method to analyse EEG data from 16 patients with disorders of consciousness and 12 healthy controls. 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 reliably responding to commands.

Goldfine et al. (2013) took issue with the statistical techniques used to detect responsiveness. 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. criticized the method of cross-validation used, the way that *p*-values were calculated, and the chosen significance threshold of $p < 0.05$. As has often been noted (e.g. Benjamin et al. 2018), 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). We should expect cases of misattributed responsiveness to arise in roughly 1 in 20 recordings. Goldfine et al. (2013) 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. (2013) replied combatively, accepting none of the criticisms. My aim is not to referee the dispute here. I will restrict myself to two comments. Firstly, 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.

4. 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. (2019), published in the *New England Journal of Medicine*. 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 shape methodological decisions.

Before going deeper into methodological details, we should note that the terms "false positive" and "false negative" can lead to confusion here. 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 (as in Cruse et al. 2011) 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 six blocks of eight trials each (i.e. 48 trials). This is called an "individualized 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 generalize 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.[2] 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" (Claassen et al. 2019, 2501).

How reliable is this result? 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?

[2] 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.

c) How will adjustments to the significance threshold be made for multiple recordings of a single patient?
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., like Cruse et al. (2011), used a standard *p*-value threshold of 0.05 to assess whether the classifier is performing significantly better than chance. A *p*-value of 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): for some but not all patients, Claassen et al. took multiple recordings. And indeed, 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 (Edlow et al. 2021). Yet doing multiple recordings drives up the chance of misattributed responsiveness.

There are two well-known statistical techniques for manage the risk: the Bonferroni procedure and the above-mentioned Benjamini-Hochberg procedure, recommended by Goldfine et al. (2013). The former is notoriously more stringent than the latter. Claassen et al. used the latter, less stringent procedure, which is less conservative and thus more tolerant of false discoveries.

To elaborate briefly on this point, the Bonferroni procedure controls the overall chance of a false positive (Type I error) 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. Yet the procedure is notoriously stringent. 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 is less stringent and aims to control a different quantity: the false discovery rate (the fraction of cases of misattributed responsiveness among the total number of positive tests). After applying this procedure, we can regard each “positive” recording, indicating apparent responsiveness, as having at most a 5% chance of being a misattribution.

The trouble here is that, in a clinical setting, it is surely the overall chance of a misattribution *for this patient* (i.e. 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. 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.

Indeed, Claassen et al. themselves note 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*" (Claassen et al. 2019, supplementary information, page 13). This implies an expected false discovery rate of 5/16, or 0.31. Note that the false discovery rate is normally defined as the fraction of all *positive* tests that are cases of misattributed responsiveness, not the fraction of *all* tests—so the expected false discovery rate is 0.31, even though the expected fraction of all tests is 0.05. The figure of 0.31 may be an underestimate because, as just noted, the control of the false discovery rate for multiple recordings of the same patient was done using the Benjamini-Hochberg procedure, when the more stringent Bonferroni procedure would have been needed to hold the absolute probability of misattributed responsiveness for each patient at 0.05 or less.

Why did Claassen et al. not control 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). Researchers in this area will tend to use a *p*-value threshold of $p < 5 \times 10^{-8}$ in order to control the false discovery rate (Chen et al. 2021). The Benjamini-Hochberg procedure, which Claassen et al. applied to multiple recordings *from the same patient*, could also have been applied to the *whole set* of tests across *all* patients. To apply it only to multiple recordings from the same patient is a choice that would be controversial in other contexts. 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.

In sum, the researchers 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, albeit 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 researchers are quite clearly, and understandably, worried about the risk of missed responsiveness. As Fins and Bernat (2018) have emphasized, the "ethical importance of avoiding type II error: failing to identify consciousness when it is present" looms large for researchers in this area. This concern drives methodological choices that prioritize avoiding cases of missed responsiveness, while expressing a more relaxed attitude towards cases of misattributed responsiveness.

5. 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. (2019) study, in which 0.31 may be an underestimate of the false discovery rate. 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.

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 (Fins 2014; Edlow and Fins 2018). 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. In some cases, a misattribution of responsiveness may give a patient's family false hope: hope that a given level of recovery is in fact realistic, when other evidence suggests it is not. 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 agonizingly 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 stabilizes, 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 (2013) have urged clinicians to take this risk seriously. Values in this area vary a great deal. Some patients would want to be kept alive even if a good recovery was very unlikely, whereas others would not (Edlow and Fins 2018).

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 (Johnson 2022). 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 recognizing this to be subject to great uncertainty (GMC-BMA 2018).

In the absence of a clear advance directive by the patient themselves, the patient's family is generally recognized to be the best (though fallible) way to access the patient's own values. And yet, one can hardly go 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 the expected false discovery rate to the patient's family? Should we say: "We expect there is a 0.31 false discovery rate associated with this procedure"? Edlow and Fins (2018) 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."

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 0.31 does not equate to a 0.31 *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—i.e. the strength of evidence varies from one case to the next—points us in the direction of a possible solution.

6. Communicating uncertainty, proposal 1: Degrees of evidence

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: (i) EEG data do not directly support a yes/no verdict on any question, but rather provide a quantitative degree of evidence; (ii) converting degrees of evidence into probabilistic assessments requires consideration of priors.

There may be a temptation to react to the high false discovery rate at $p < 0.05$ 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 $p < 0.005$ (Benjamin et al. 2018). But 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? Here is a first step. 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. The relationship between *p*-values and strength of evidence is, admittedly, 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 dataset 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 (Benjamin and Berger 2019). 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 (2019) recommend the increased use of BFBs in science as a response to the replication crisis. 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. Of course, two problems arise: most human clinicians will not know how to convert a *p*-value to a BFB, nor will they know how to move from a BFB to a probability that the patient is responding. The first problem is easily remedied, since the AI product itself could

calculate a BFB. The second problem is more serious, because prior probabilities are needed to generate posterior probabilities from Bayes factors.

7. 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:

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

But how to set these priors? One option would be to set our priors in line with base rates. At present, our uncertainty about individual cases of cognitive-motor dissociation 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 (i.e. 16/104, minus 5 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 (e.g. 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 background information that might relevantly shape the priors, 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 call for expert judgement.

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 ratio (see Section 8), so it will be helpful if the formula also converts the odds to a probability by applying the formula:

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

Yet this proposal comes with an associated risk: a risk of a probabilistic variant of the "self-fulfilling prophecy", whereby a clinician inputs priors that give extremely low (or, less plausibly, extremely high) odds of responsiveness, making it very difficult for even strong evidence to shift the odds. And recall the background: 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 realized.

This brings us to the nub of the problem: how 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 conscious states that clinicians are unable to detect without technological assistance. My proposed solution is that clinicians should be advised to input priors that fall within an *appropriately cautious* range, after discussion with all parties involved in the clinical care of the patient, *including the patient's family*. These priors should be non-extreme, non-dogmatic, and heavily influenced by recommendations from professional bodies.

For this proposal to be implemented in practice, clinicians would need to have access to clear guidance from professional bodies. 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. 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.[3]

8. 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?

One problem here is that, without standardized language, numerical probabilities can be coverted to ordinary language terms in many different ways (Lipkus 2007). A "probabilistic yardstick" aims to solve this problem by providing a standardized protocol for assigning verbal, qualitative labels to probability ranges. An influential example is the PHIA (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 vaguely bounded ranges of probabilities.

[3] Merker (2007); Panksepp (1998, 2005, 2011); Panksepp et al. (2007).

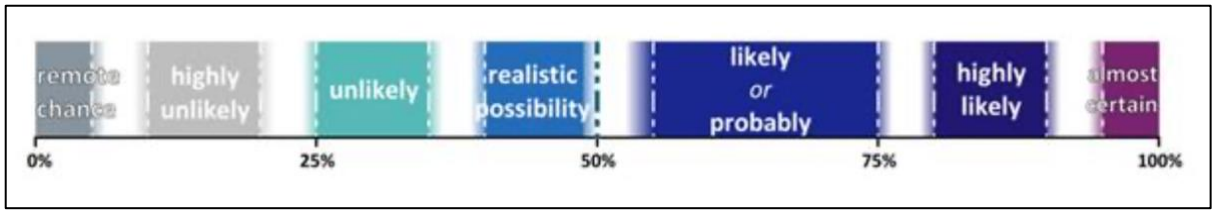

Figure 1. *The PHIA probability yardstick. This figure is reproduced from SPI-M-O (2022) but the same figure can be found widely in public domain UK government documents.*

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 standardized 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 (Birch et al. 2022; Ongena et al. 2021). 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 (Lipkus 2007). 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 <u>at most</u> 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.*

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.

9. 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 own 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 upper bounds; (2) the use of *patient-centred priors* to convert degrees of evidence to probabilistic assessments; and (3) the use of standardized *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 (Birch et al. 2022). 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 that the patient would not want to be made. 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 (Birch et al. 2022; Ongena et al. 2021). 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 (Johnson 2022). 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.

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

References

Bai, Yang, Yajun Lin and Ulf Ziemann. 2021. "Managing Disorders of Consciousness: The Role of Electroencephalography." *Journal of Neurology* 268:4033-4065. https://doi.org/10.1007/s00415-020-10095-z

Bavli, Itai and Daniel Steel. 2020. "Inductive Risk and OxyContin: The Ethics of Evidence and Post-Market Surveillance of Pharmaceuticals in Canada." *Public Health Ethics* 13(3):300-313.

Bayne, Tim, Jakob Hohwy and Adrian M. Owen. 2016. "Are There Levels of Consciousness?" *Trends in Cognitive Sciences* 20(6):405-413. http://dx.doi.org/10.1016/j.tics.2016.03.009

Benjamin, Daniel J. and James O. Berger. 2019. "Three Recommendations for Improving the Use of *p*-Values." *The American Statistician* 73(supplement 1): 186-191, https://doi.org/10.1080/00031305.2018.1543135

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, *et al.* 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2:6–10. https://doi.org/10.1038/s41562-017-0189-z.

Biddle, Justin B. 2016. "Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease." *Perspectives on Science* 24 (2):192-205.

Birch, Jonathan, Kathleen A. Creel, Abhinav K. Jha and Anya Plutynski. 2022. "Clinical Decisions Using AI Must Consider Patient Values." *Nature Medicine* 28:229-232. https://doi.org/10.1038/s41591-021-01624-y

Bluhm, Robyn. 2017. "Inductive Risk and the Role of Values in Clinical Trials." In Kevin Elliot & Ted Richards (eds.), *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, pp. 193-214.

Claassen, Jan, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M. Burger, Angela Velazquez, Joshua U. Okonkwo, Jean-Rémi King, Soojin Park, Sachin Agarwal, David Roh, Murad Megjhani, Andrey Eliseyev, E. Sander Connolly, and Benjamin Rohaut. 2019. "Detection of Brain Activation in Unresponsive Patients with Acute Brain Injury." *New England Journal of Medicine* 380(26):2497-2505. https://doi.org/10.1056/NEJMoa1812757

Cruse, Damian, Srivas Chennu, Camille Chatelle, Tristan A. Bekinschtein, Davinia Fernández-Espejo, John D. Pickard, Steven Laureys and Adrian M. Owen. 2011.

"Bedside Detection of Awareness in the Vegetative State: A Cohort Study." *Lancet* 378(9809):2088–2094. https://doi.org/10.1016/S0140-6736(11)61224-5

Cruse, Damian, Srivas Chennu, Camille Chatelle, Tristan A. Bekinschtein, Davinia Fernández-Espejo, John D. Pickard, Steven Laureys and Adrian M. Owen. 2013. "Reanalysis of 'Bedside detection of awareness in the vegetative state: a cohort study' – Authors' reply." *Lancet* 381(9863):P291-P292. https://doi.org/10.1016/S0140-6736(13)60126-9

Douglas, Heather E. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Edlow, Brian L. and Joseph J. Fins. 2018. "Assessment of Covert Consciousness in the Intensive Care Unit: Clinical and Ethical Considerations." *Journal of Head Trauma Rehabilitation* 33(6) 424–434. https://doi.org/10.1097/HTR.0000000000000448

Edlow, Brian L., Jan Claassen, Nicholas D. Schiff and David M. Greer. 2021. "Recovery from Disorders of Consciousness: Mechanisms, Prognosis and Emerging Therapies." *Nature Reviews Neurology* 17:135-156. https://doi.org/10.1038/s41582-020-00428-x

Fins, Joseph J. 2014. *Rights Come to Mind: Brain Injury, Ethics, and the Struggle for Consciousness.* Cambridge: Cambridge University Press.

Fins, Joseph J. 2013. "Disorders of Consciousness and Disordered Care: Families, Caregivers, and Narratives of Necessity." *Archives of Physical Medicine and Rehabilitation* 94(10):1934–1939

Fins, Joseph J. and James L. Bernat. 2018. "Ethical, Palliative, and Policy Considerations in Disorders of Consciousness." *Neurology* 91:471-475. https://doi.org/10.1212/WNL.0000000000005927

Royal College of Physicians and British Medical Association [RCP-BMA]. 2018. *Clinically-Assisted Nutrition and Hydration (CANH) and Adults Who Lack the Capacity to Consent: Guidance for Decision-Making in England and Wales.* https://www.bma.org.uk/canh

Goldfine, Andrew M., Jonathan C. Bardin, Quentin Noirhomme, Joseph J. Fins, Nicholas D. Schiff and Jonathan D. Victor. 2013. "Reanalysis of 'Bedside Detection of Awareness in the Vegetative State: A Cohort Study.'" *Lancet* 381(9863):289–291. http://doi.org/10.1016/S0140-6736(13)60125-7

Johnson, L. Syd. M. 2022. *The Ethics of Uncertainty: Entangled Ethical and Epistemic Risks in Disorders of Consciousness.* New York: Oxford University Press.

Karaca, Koray (2021). “Values and Inductive Risk in Machine Learning Modelling: The Case of Binary Classification Models.” *European Journal for Philosophy of Science* 11(4):1-27.

Kondziella, D., A. Bender, K. Diserens, W. van Erp, A. Estraneo, R. Formisano, S. Laureys, L. Naccache, S. Ozturk, B. Rohaut, J. D. Sitt, Stender, J., Tiainen, M., Rossetti, A.O., Gosseries, O., Chatelle, C. 2020. European Academy of Neurology guideline on the diagnosis of coma and other disorders of consciousness. *European Journal of Neurology* 27:741-756. https://doi.org/10.1111/ene.14151

Kitzinger, Jenny and Celia Kitzinger. 2013. “The ‘Window of Opportunity’ for Death After Severe Brain Injury: Family Experiences.” *Sociology of Health & Illness* 35(7):1095–1112. https://doi.org/10.1111/1467-9566.12020

Kitzinger, Celia and Jenny Kitzinger. 2015. “Withdrawing Artificial Nutrition and Hydration from Minimally Conscious and Vegetative Patients: Family Perspectives.” *Journal of Medical Ethics* 41(2):157–160.

Kitzinger, Jenny and Celia Kitzinger. 2018. “Deaths after Feeding-Tube Withdrawal from Patients in Vegetative and Minimally Conscious States: A Qualitative Study of Family Experience.” *Palliative Medicine* 32(7):1180–1188.

Kostko, Aaron. 2019. “Inductive Risks and Psychiatric Classification”. In Serife Tekin and Robyn Bluhm (eds.), *The Bloomsbury Companion to Philosophy of Psychiatry*. London: Bloomsbury, pp. 197-216. http://dx.doi.org/10.5040/9781350024090.ch-010

Kouider, Sid, Thomas Andrillon, Leonardo S. Barbosa, Louise Goupil and Tristan A. Bekinschtein. 2014. “Inducing Task-Relevant Responses to Speech in the Sleeping Brain.” *Current Biology* 24:2208-2214. http://dx.doi.org/10.1016/j.cub.2014.08.016

Kukla, Rebecca. 2019. “Infertility, Epistemic Risk, and Disease Definitions.” *Synthese* 196(11):4409-4428.

Lewens, Tim. 2019. “The Division of Advisory Labour: The Case of ‘Mitochondrial Donation’”. *European Journal for Philosophy of Science* 9(1):1-24.

Lipkus, Isaac M. 2007. “Numeric, Verbal, and Visual Formats of Conveying Health Risks: Suggested Best Practices and Future Recommendations.” *Medical Decision Making* 27:696-713. https://doi.org/10.1177/0272989X07307271

Merker, Bjorn. 2007. “Consciousness Without a Cerebral Cortex: A Challenge for Neuroscience and Medicine.” *Behavioural and Brain Sciences* 30:63-81. https://doi.org/10.1017/S0140525X07000891

Ongena, Yfke P., Derya Yakar, Marieke Haan and Thomas C. Kwee. 2021. “Artificial Intelligence in Screening Mammography: A Population Survey of Women's Preferences.” *Journal of the American College of Radiology* 18(1):P79–P86. https://doi.org/10.1016/j.jacr.2020.09.042

Panksepp, Jaak. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Panksepp, Jaak. 2005. “Affective Consciousness: Core Emotional Feelings in Animals and Humans.” *Consciousness and Cognition* 14:30-80. https://doi.org/10.1016/j.concog.2004.10.004

Panksepp, Jaak. 2011. “The Basic Emotional Circuits of Mammalian Brains: Do Animals Have Affective Lives?” *Neuroscience and Biobehavioral Reviews* 35:1791-1804. https://doi.org/10.1016/j.neubiorev.2011.08.003

Panksepp, Jaak, Thomas Fuchs, Victor Abella Garcia and Adam Lesiak. 2007. “Does Any Aspect of Mind Survive Brain Damage that Typically Leads to a Persistent Vegetative State? Ethical Considerations.” *Philosophy, Ethics, and Humanities in Medicine* 2:32. https://doi.org/10.1186/1747-5341-2-32

Peterson, Andrew, Damian Cruse, Lorina Naci, Charles Weijer and Adrian M. Owen. 2015. “Risk, Diagnostic Error, and the Clinical Science of Consciousness.” *NeuroImage: Clinical* 7:588-597. https://dx.doi.org/10.1016/j.nicl.2015.02.008

Peterson, Andrew, Adrian M. Owen and Jason Karlawish. 2020. “Translating the Discovery of Covert Consciousness into Clinical Practice”. *JAMA Neurology* 77(5):541-542. https://doi.org/10.1001/jamaneurol.2020.0232

Plutynski, Anya. 2017. “Safe, or Sorry? Cancer Screening and Inductive Risk.” In Kevin C. Elliott & Ted Richards (eds.), *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, pp. 149-169.

Scarantino, Andrea. 2010. “Inductive Risk and Justice in Kidney Allocation.” *Bioethics* 24(8):421-430.

Scientific Pandemic Influenza subgroup on Modelling (SPI-M-O). 2022. “SPI-M-O Consensus Statement on COVID-19, 6 January 2022.” https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1046467/S1477_SPI-M-O_consensus_statement.pdf

Stanev, Roger. 2017. “Inductive Risk and Values in Composite Outcome Measures.” In Kevin Elliot & Ted Richards (eds.), *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, pp. 171-192.

Stegenga, Jacob. 2017. “Drug Regulation and the Inductive Risk Calculus.” In Kevin C. Elliott & Ted Richards (eds.), *Exploring Inductive Risk: Case Studies of Values in Science*. New York: Oxford University Press, pp. 17-36.

Turgeon, Alexis F., François Lauzier, Jean-François Simard, Damon C Scales, Karen E. A Burns, Lynne Moore, David A. Zygun, Francis Bernard, Maureen O. Meade, Tran Cong Dung, Mohana Ratnapalan, Stephanie Todd, John Harlock, Dean A. Fergusson, Canadian Critical Care Trials Group. 2011. “Mortality Associated with Withdrawal of Life-Sustaining Therapy for Patients with Severe Traumatic Brain Injury: A Canadian Multicentre Cohort Study.” *Canadian Medical Association Journal* 183(14):1581-8. https://doi.org/10.1503/cmaj.101786