Received: 5 April 2019 | Revised: 28 October 2019 | Accepted: 19 November 2019

DOI: 10.1002/hbm.24886

WILEY

RESEARCH ARTICLE

Your evidence? Machine learning algorithms for medical diagnosis and prediction

Bert Heinrichs[1,2] | Simon B. Eickhoff[3,4]

[1]Institute of Neurosciences and Medicine, Ethics in the Neurosciences (INM-8), Research Center Jülich, Jülich, Germany

[2]Institute of Science and Ethics (IWE), University of Bonn, Bonn, Germany

[3]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[4]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Germany

Correspondence
Bert Heinrichs, Institute of Neurosciences and Medicine, Ethics in the Neurosciences (INM-8), Research Center Jülich, Wilhelm-Johnen-Straße, 52425 Jülich, Germany.
Email: b.heinrichs@fz-juelich.de

Abstract

Computer systems for medical diagnosis based on machine learning are not mere science fiction. Despite undisputed potential benefits, such systems may also raise problems. Two (interconnected) issues are particularly significant from an ethical point of view: The first issue is that epistemic opacity is at odds with a common desire for understanding and potentially undermines information rights. The second (related) issue concerns the assignment of responsibility in cases of failure. The core of the two issues seems to be that understanding and responsibility are concepts that are intrinsically tied to the discursive practice of giving and asking for reasons. The challenge is to find ways to make the outcomes of machine learning algorithms compatible with our discursive practice. This comes down to the claim that we should try to integrate discursive elements into machine learning algorithms. Under the title of "explainable AI" initiatives heading in this direction are already under way. Extensive research in this field is needed for finding adequate solutions.

KEYWORDS

discursive practice, epistemic opacity, explainability, machine learning, medical diagnosis, medical ethics, medical prediction, responsibility, understanding

1 | INTRODUCTION

There is a famous scene in the movie "Harry Potter and the Half-Blood Prince": A student has been cursed, investigations are under way. Professor McGonagall and Professor Snape question Harry and his friends Hermione and Ron about the incident. All at once, Harry shouts "It was Malfoy." McGonagall replies "This is a very serious accusation, Potter." "Indeed," agrees Snape and continues "Your evidence?" Harry immediately responds, "I just know." Snape—superbly played by the late Alan Rickman—retorts "You just… know." As spectators we are, of course, on Harry's side. However, we feel that his answer is not quite convincing. He should provide something more substantial to accuse his arch-rival Malfoy. Simply knowing is not sufficient—even for Harry Potter.

The same holds, even more, for ordinary people in situations of everyday life, in particular if stakes are high. An important case in point is medical practice. If a physician makes a diagnosis or recommends a treatment and her patient asks for evidence or an explanation, the assurance "I simply know" is hardly appropriate to settle the case. Very likely, the patient would leave the doctor's office and look out for a more skilled or more communicative colleague. However, what if the physician were to use an automated system based on artificial intelligence, which analyses all the available data of the patient and, based on a previously trained algorithm, suggests the diagnosis? Of course, in this case the physician could declare, “I know because the computer told me so.” The patient could continue to ask, “How does the computer know?” At this point, the only answer available to a physician might be “It simply knows.” At least, that would be the only available answer if the physician had access to nothing more than the model's output.

In this article, we aim at addressing some critical issues raised by the use of machine learning algorithms for medical diagnosis and prediction. We start with examining the notion of interpretability and how it is related to machine learning. Then, we give a brief overview of the state of the art in medical AI. Against this background, we put forward what we consider two crucial issues: The first issue is that epistemic opacity is at odds with a common desire of understanding and potentially undermines information rights. The second (related) issue concerns the assignment of responsibility in cases of failure. Subsequently, we elaborate these issues in detail. Thereafter, we suggest that explainable AI might help to overcome some of the problems. Finally, we look at some of the implications for medical practice in general and for neuroimaging in particular.

2 | THE ISSUE OF INTERPRETABILITY

The reason for this above mentioned predicament is that most current machine learning algorithms give rise to what has been called “the black-box problem” (Castelvecchi, 2016). More specifically, most machine learning algorithms are characterized, albeit to a varying degree, by “epistemic opacity.” That is to say, these algorithms include epistemically relevant elements, which a cognitive agent cannot (easily) access (Humphreys, 2009). It is important to note, although, that access may be pursued by two distinct yet interacting routes, namely model and results interpretability.

Results interpretability will eventually be critical in a medical context, as it resembles the information asked for in the described scenario. It consists of information detailing what aspects of the data have led to a certain decision in a particular case. In this it is akin to what a physician traditionally tells a patient, for example, “Given your age, your heart is slightly enlarged. Furthermore, you have a history of moderate asthma. On the other hand, your overall constitution is good. Experience shows that in female patients this combination is rather common and is, in most cases, not alarming. Presently, there is no need to take further action. You should check again in six months.” A model detailing the reasons for a particular decision on an individual case in plain language like this is presently unavailable and will probably remain an unmet challenge for quite some time.

Model interpretability refers to a human's ability to understand the model itself or at least a summary thereof. That is, the distinction between results versus model interpretability can be conceptualized as understanding individual decisions versus understanding the model, respectively. The latter than entails to know which input features are actually used (selected) by the model, how these features are weighted and combined, and how decisions are derived from this process. There are also critical distinctions in the role results and model interpretability will play in clinical application. Explainable results (cf. above) can be directly used to integrate the recommendation by an algorithm into clinical decisions and to communicate the reason for a particular suggestion to the patient. In contrast, model interpretability provides more generic, “background” knowledge for shared decision-making. In particular, only an interpretable model, that is, one that allows humans to gain knowledge about the features that are considered, their integration and weighting puts the physician in a position where she can process and interpret the results of the algorithm relative to information from various other sources and her individual evaluation of the case. That is, information on the way the model works are relevant for connecting it to other information such as clinical history or lab-results and forming a final (clinical) decision that can be communicated and explained to the patient. At the same time, model interpretability is crucial for estimating the plausibility of a result generated by a machine learning algorithm. Importantly, model interpretability depends strongly on the type of learning algorithm employed, with a general trade-off between performance and interpretability reflecting model complexity as detailed below. Critical distinctions in this context may be made between sparse models that are trained to perform as well as possible on a minimal number of input features and those that do not have this constraint, as well as between models based on a single training and those employing ensemble methods, that is, those which are trained repeatedly on subsets of the training sample. In general, sparse models are more interpretable, and ensemble models are less so.

One of the key reasons that interpretability has become a more prominent issue over the recent years, even though techniques such as support vector machines or decision trees have been in use since the 1990s, is the emergence of “deep learning” over the past decade. Advances in both algorithms (Hinton, 2007) and hardware (GPUs and subsequently TPUs) have resulted in much more efficient and powerful estimation of deep architectures, and have subsequently given rise to a distinction featuring prominently in current debates, namely the one between “conventional machine learning” and “deep learning.” According to LeCun, Bengio and Hinton conventional machine-learning techniques are “limited in their ability to process natural data in their raw form,” while deep learning methods “are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into representation at a higher, slightly more abstract level. […] The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure.” (LeCun, Bengio, & Hinton, 2015, p. 436) This means that while deep learning algorithms are trained on big data collections they develop in a way that is neither foreseeable nor transparent to the programmer. In the present context, two aspects deserve special attention. First, because deep learning does not rely on feature engineering but rather on high-dimensional representations in multi-layer networks, it has been argued that network engineering, that is, the art of constructing layers and functions performed by these, has effectively taken the same place that feature engineering previously occupied (Maier et al., 2018). Second, by means of combination of complex ensembles, in particular those containing nonlinearities, a similar situation as just described for deep learning may arise for classical machine-learning approaches. In short, different types of algorithms may perform at a level that is comparable to the best clinical practice of experienced physicians. Yet, it may be unclear why and how they do so.

3 | STATE OF THE ART IN MEDICAL AI

Computer systems for medical diagnosis based on machine learning are not mere science fiction (Kononenko, 2001). One illustrative example is a recent study, which reported that deep convolutional neural networks are able to classify skin lesions achieving a level of performance on par with experts in view of the identification of the most common cancers as well as regarding the identification of the deadliest skin cancer (Esteva et al., 2017). The authors saw this as "demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists." (Esteva et al., 2017, p. 115). In view of this (and similar) claims it needs to be taken into account that almost all applications are related to supervised problems, that is, the algorithms are provided with labeled training data and have the task of extracting patterns from the data that enables the prediction of the respective labels in new subjects. However, the training labels (diagnoses for the cases that were used to learn the association between features and outcomes) and ground truth for the evaluations (diagnoses for the cases that were used to assess the performance of the algorithm by putting their features into the trained model and comparing the prediction to the known diagnoses) are traditional clinical diagnoses which are noisy and in some cases highly disputable. Talking about "close to perfect" or even "super-human" performance is, therefore, somewhat misleading. In any case, it leads to the interesting situation that the performance of a theoretically perfect AI that recognizes a real structure may actually be worse as compared to an algorithm that, like most current approaches, attempts to re-create clinical diagnoses (Arbabshirani, Plis, Sui, & Calhoun, 2017).

In other areas of medicine similar systems are the object of research. Related methods have, for example, been applied to decipher the data from stroke imaging and have demonstrated some promising results (Eun-Jae, Yong-Hwan, Namkug, & Dong-Wha, 2017). Additionally, machine-learning frameworks for early MRI-based Alzheimer's conversion prediction in MCI subjects are being explored (Moradi et al., 2015). Moreover, the possibility of making individual prognoses in psychiatry using neuroimaging and machine learning is under investigation (Arbabshirani et al., 2017; Bzdok & Meyer-Lindenberg, 2018; Janssen, Mourao-Miranda, & Schnack, 2018).

We note that the challenges discussed in this paper should hold rather generically for all current and most likely future approaches to using machine learning for medical decision-making, as they point to fundamental issues in clinical translation and acceptance. Yet, there is evidently a broad spectrum of conceptually and in particular technically different methods that are currently being used in medical applications including imaging neuroscience. For a more detailed overview, we would like to point the reader to other, more specialized publications (Bishop, 2006; Bzdok & Ioannidis, 2019; Choudhary & Gianey, 2017; James, Witten, Hastie, & Tibshirani, 2013; Jordan & Mitchell, 2015). What is critical to note in the current context, though, is that there is generally an inverse relation between the potential accuracy or performance of machine-learning algorithms on one hand and their interpretability on the other. That is, approaches that trend to be very good in predicting new cases trend to be least transparent and hence do allow little insight into the evidence for this decision. While the exact order can be discussed (cf. references above), the major classes of machine-learning algorithms can be roughly aligned on this spectrum as follows.

Models based on linear regression, including regularized approaches such as LASSO, Ridge Regression or linear Support Vector Machines are most interpretable, since their decision value is a linear combination of the features. Decision trees are often conceived to be on a similar level along the accuracy/transparency trade-off. While they allow for more complex, nonlinear interactions between features through a sequence of splits, they are still rather well interpretable for humans even though inspection can become challenging for deeper trees. K-nearest neighbor algorithms provide an example for instance-based or "lazy" learning in which the input is compared to the training datasets and receives the label that was most frequent among those training observations that were most similar to it. While these trend to outperform the previously mentioned models, interpretation is often complicated by the fact that "similarity" in high-dimensional spaces is itself hard to describe from the human perspective. That is, while interpretation could be possible by comparison to the selected neighbors, why these instances turned out to be the closest neighbors in high-dimensional space is often harder to appreciate. The same holds true for kernel regression, which can be conceptualized as a weighted form of the previous approach, that is, each training observation contributes to the prediction according to its similarity with the test sample. As in this case many, potentially all, training observations contribute, interpretability is usually lower than for nearest neighbor approaches. Ensemble based approaches, which repeatedly fit simpler models (such as decision trees or linear models) on parts of the data and then combine their results into a final prediction have enjoyed considerable popularity and success in recent years, as these trend to show very good accuracies even in moderately sized training sets and can be optimized to avoid unbiases through, for example, stratified subsampling, their interpretability is severely limited. Finally, deep neural networks have been shown to provide the best accuracy in many fields, but also represent the class of tools, which has the lowest transparency, as the decisions are made through nonlinear interactions involving the optimization of often millions of parameters.

In general, classes of models that tend to yield good accuracies and should hence be most interesting from the application perspective come with the challenge of being complex and hence not readily interpreted by humans. We would like to note, though, that the postulated order only holds in the case of sufficiently large training samples, that is, when performance is limited by the model, not the data. This, however, is often not the case in medical applications where the number of observations (patients) is often limited. This may explain the good performance and hence considerable success "simple" models like SVM enjoy in the context of translational machine learning, for example, in neuroimaging (He et al., 2018).

Despite undisputed potential benefits, systems for medical diagnosis and prediction based on machine-learning algorithm, in particular those involving ensembles, nonlinear mapping, and deep neural networks, may also raise problems. Two (interconnected) issues are particularly significant from a philosophical point of view: The first issue is that epistemic opacity is at odds with a common desire of understanding and potentially undermines information rights. The second (related) issue concerns the assignment of responsibility in cases of failure.

4 | ASKING FOR EVIDENCE

Just like in the case of Professor Snape and Harry Potter, we regularly ask others for evidence when they make an assertion. One apparent reason for doing this is that we try to estimate the degree of certainty or confidence that we should assign to the assertion in question. If Harry had answered "I saw Malfoy waving his wand and casting a spell", Snape would probably have counted this as rather strong evidence. If, in contrast, he had replied "Gini told me that Fred and George heard that a third-year student from Hufflepuff suspected that it must have been Malfoy", nobody else would have been convinced. In other words, by asking for evidence we try to find out whether an assertion is true. If there is strong evidence in favor of it, we have good reasons to take it to be true; if evidence is less convincing, we classify an assertion as a mere conjecture.

However, investigating certainty or confidence is not all we do when we ask for evidence. Even if we are convinced that another person is highly reliable we will still ask for evidence. If so, we try to make the assertion intelligible for us. The same holds for the results generated by a machine-learning algorithm where the reliability of a human corresponds to good performance in a priori evaluations. In contrast to quantified (numerical) confidence ratings which simply have to be taken for granted, further evidence allows for a more comprehensive decision-making process. In this process, human agents can bring in their own experiences, which makes it possible to consider recommendations against the background of individual attitudes. Of course, this process caters to the psychological bias of self-centrism that makes us trust our own judgments more than quantitative confidence ratings, even if it leads to suboptimal results. Aside from the self-serving psychological effects including a sense of agency and positive reinforcement, a key contribution of this process is the inclusion of previous knowledge of the "receiver" (i.e., the patient or, in the case of AI based recommendations, the physician) into a coherent set of beliefs.

In contrast, as long as supporting evidence of this sort is lacking, an assertion will inevitably remain isolated and we will find it difficult to integrate it into our broader "web of beliefs" (Quine & Ullian, 1978). This is not only relevant for the assessment of diagnoses, but probably even more so for decisions about the consequences, for example, about which treatment to choose. Even if we have the strong feeling that an assertion might be true, lack of evidence makes it suspicious. For we are often not, or at least not only, interested in knowledge but in understanding. High confidence assessments from reliable agents, which machine-learning will most likely provide in the future, may be under-valued by human decision makers if they are opaque and unmatched to personal experiences.

Traditionally, epistemology has focused on knowledge. Knowledge is the central concept which epistemology tries to elucidate, probably even more since Gettier showed that the well-known analysis of knowledge as justified true belief is flawed or at least incomplete (Gettier, 1963). Until recently, epistemology paid less attention to other epistemic states, in particular to understanding (Kvanvig, 2003). This is somewhat surprising since it seems initially plausible that understanding is epistemically equally, if not more valuable than knowledge or justified true belief. Now, there is an ongoing controversy among epistemologists what exactly understanding is, what its features are and why its value is particularly high (Bondy, 2015). For the present purposes, a proposal by Christoph Kelp is particularly helpful. Criticizing "explanationist views" on the one side and "manipulationist views" on the other side, he has suggested an account of understanding as "well-connected knowledge" (Kelp, 2015). According to this view, understanding a phenomenon P does not only involve knowing a set of true propositions p_i about P, but also knowing how these propositions are interrelated. Notably, on this account "understanding" is a concept that allows for degrees (Kelp, 2015, pp. 3809–3813). For example, one understands P better if one knows, in addition to p_1 to p_4 being true, that p_1 and p_2 entail p_3 and that p_4 gives a causal explanation of p_1 in terms of p_3 (Kelp's concrete example being the reaction of litmus paper to the application of acidic substances).

The important point with regard to most machine-learning algorithms is that *by design* they render understanding in the sense of well-connected knowledge impossible. Their internal development is inevitably opaque, irrespective of recent attempts and advances such as saliency maps of feature importance since the information on which aspects of the data contributed to the decision usually does not allow a description of how they contribute and how a decision was reached through their interaction. If, for example, a machine-learning algorithm indicates that a patient is at high risk for developing Parkinson's (= p_1), it remains unclear how this is exactly related to other available medical data, for example, MRI scans (= p_2), lab results (= p_3) and the clinical history (= p_4). Simply knowing p_1 to p_4 does not imply understanding the patient's condition. Even in the case that p_2 to p_4 support the prediction, the logical relation of p_1 to p_2, p_3, and p_4, respectively remains unclear. In contrast, the relation between p_2, p_3, and p_4 is accessible. It might, for example, be that the clinical history (p_4) on the one side and the MRI scans and lab results (p_2 and p_3)

stand in an implication relation (p_4→p_2 and p_4→p_3). Knowing this substantially contributes to the overall understanding of p_2, p_3, and p_4. The ability of machine-learning approaches to derive complex patterns of high-dimensional interactions lies at the heart of their power and renders the process opaque. The outcome they generate does not allow for connecting them with previous knowledge. What we get from them is, at best, knowledge or justified true beliefs but no understanding. However, as has been argued before, in many situations it is exactly understanding that we are looking for. In particular, in the case of medical diagnosis and prediction many people ask for evidence in this later sense. Physicians will need it to take the responsibility for the final diagnosis. Patients often wish for it to be able to accept it and in some cases to make it psychologically bearable.

Of course, this does not hold for all patients and under all circumstances. Frequently, what patients are seeking is confidence in the method employed. Let us assume, a physician suggests a blood test for detecting a specific disease marker. Most patients are pleased with grasping why the physician recommends it rather than understanding how exactly the test works and how the particular marker is related to the pathophysiology of his disorder. Moreover, some patients would even be annoyed if the physician started to elaborate on scientific details. However, even those patients would probably find it irritating to learn that the physician has no clue about how the blood test works and why the results are relevant to managing their particular disease. For them, understanding is important, but not "first-person" understanding. Rather, by being assured about the knowledge of the physician and their confidence in the approach, the patients will become content about being treated well. Furthermore, there undoubtedly are patients for whom "first-person" understanding is vital, in particular in view of machine-learning algorithms. This is, at least, suggested by the new EU General Data Protection Regulation (GDPR) as of 2016, which became effective in May 2018. Recital 71 of the GDPR reads as follows: "The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. […] In any case, such processing should be subject to suitable safeguards, which should include *specific information to the data subject* and the right to obtain human intervention, to express his or her point of view, *to obtain an explanation of the decision reached* after such assessment and to challenge the decision." (Council Regulation, 2016 emphasis added) Even when considering the inevitable need for further weighting the right for explanation with other goods (like the need for rapid intervention in emergencies), this provision can be taken as an indicator that explanation and understanding are important for many people—either first-person or by proxy, for example, by a physician who is involved in the decision-making. Some legal scholars (e.g., Selbst & Powles, 2017) even argue that the GDPR includes a "right to an explanation" while others (e.g., Wachter, Mittelstedt, & Floridi, 2017) reject such a reading as too strong. At any rate, the fact that the EU has decided to set the GDPR into force suggests that automated processing is a major concern of many people.

5 | ASSIGNING RESPONSIBILITY

Imagine that the medical diagnosis of a system based on machine learning turned out to be mistaken. Although reliable in the past, and having achieved excellent performance in independent validations, it presented a wrong result this time. Notably, this scenario is not just possible, it will happen inevitably, as no system can achieve 100% accuracy in the real world. The drugs that were prescribed by the physician on the basis of the diagnosis provided by the algorithm did not show the intended effect and the situation of the patient declined rapidly. He comes to see the doctor again and accuses her of mistreatment. To defend herself, the doctor would normally recapitulate the evidence that initially justified her conclusion. Other medical experts could review this excuse, compare it to their own experience and established guidelines, and see whether she negligently missed or misjudged something. Eventually, the degree of responsibility for the mistreatment that we attribute to the doctor depends on the nature of the evidence that she overlooked or misinterpreted. If she missed something obvious or drew conclusions that are clearly at odds with established medical knowledge we would classify her behavior as grossly negligent. If, however, literally no expert would have noticed the crucial point or if the case represents one of the inevitable outliers of the general rule, we would probably absolve her from guilt (Figure 1).

The second issue of diagnosis based on machine learning is that it may undermine exactly this process. On the one hand, quantification of confidence as outlined above will provide numerical information on how likely a certain diagnosis or outcome is to be expected relative to a large training dataset. That is, experts' opinion on whether a failure was foreseeable or not will need to be reconciled with this quantification, allowing for conflicting interpretations. Would the AI-based diagnosis that the physician followed trump the consensus expert opinion to the contrary and hence make her nonresponsible? If so, would not that mean that the algorithm needs to take responsibility for the (wrong) suggestion? This is undoubtedly a rather problematic implication. In any case, such a judgment would only be reasonable if interpretability is ensured. In view of a "black box" we can, in principle, not recapitulate evidence employed within "the box" and, as a consequence, not evaluate the degree of responsibility. Importantly, this is not necessarily the fault of the programmers of the system. While it is true that they initially implemented certain success functions, during the learning process the algorithm turned into something new, something epistemically opaque or "inscrutable" (for more details on the latter notion see Selbst & Barocas, 2018, pp. 1094–1096). As a consequence, the assignment of responsibility is no longer possible—at least in the conventional way in which we assign responsibility to individuals. There is an ongoing debate the notion of responsibility and its applicability to autonomous systems. Notably, Matthias (Matthias, 2004) has identified a "responsibility gap" that, according to him, cannot be bridged by the received concepts of responsibility. Recently,

FIGURE 1 Illustration of our (conceptual) argument

Nyholm (2018) has argued that the right way to understand the agency of autonomous systems is in terms of human-machine collaborations. As a consequence, he sees little room for a responsibility gap. In our view, the problem of assigning responsibility provides strong reasons for the view that “medical AI” will not replace physicians in even the more distant figure. Rather, it seems most plausible that machine learning and image analytics will take a similar role as laboratory examinations have today: Providing quantitative assessments that are integrated with each other and impressions from personal examination, weighted, and combined into a final assessment. However, as noted before, integration becomes possible only through interpretability. Arguably, integration in a very weak sense is possible without interpretability—exactly by referring to reliable outcomes of an algorithm in the past. However, it is difficult to see how this alone would allow for weighing and combining a result with other opposing results (e.g., from traditional clinical diagnosis). This, in turn, refers back to the problem of responsibility. Taking responsibility includes balancing contradicting results against each other and inferentially authorizing the final decision.

6 | DISCURSIVE PRACTICE AND INFERENTIAL VAGRANTS

The core of the aforementioned issues seems to be the following: Understanding and responsibility are concepts that are intrinsically tied to the discursive practice of giving and asking for reasons. In fact, by making an assertion we place it into the inferentially structured space of reasons and at the same time take responsibility for its being true. As Robert Brandom puts it: “Saying or thinking *that* things are thus-and-so is undertaking a distinctive kind of *inferentially* articulated commitment: putting it forward as a fit premise for further inferences, that is, *authorizing* its use as such a premise, and undertaking *responsibility* to entitle oneself to that commitment, to vindicate one's authority, under suitable circumstance, paradigmatically by exhibiting it as the conclusion of an inference from other such commitments to which one is or can become entitled.” (Brandom, 2000, p. 11) This is exactly what Professor Snape calls on Harry to do by asking what the evidence of his accusation is. He calls on Harry to be (morally) accountable for his claim. Without further evidence, the accusation is almost meaningless just as a medical diagnosis or prediction would be almost meaningless without a look at the evidence, that is, the way in which the decision was reached based on the available features. Yet, even state-of-the-art algorithms are not capable of taking responsibility and due to their inherent opacity, neither is the physician who uses it, nor is the programmer who developed it. After the initial training phase, one may want to ascribe a certain form of “intelligent behavior” to such an algorithm, since it produces expedient results to complex problems, but its outcome does not qualify for a true commitment and in fact is fundamentally different from human intelligence. Moreover, as a kind of pseudocommitment, the outcome undermines our discursive practice by generating something that, at first sight, appears as a premise for further inferences. However, since no one can take responsibility for the claim, it is an inferential vagrant. To be sure, the reliability of an algorithm can be considered as a kind of connecting factor, which allows for a responsible handling of its outcomes. However, in the case of contradicting results from other sources a lack of interpretability renders it impossible to evaluate its "inferential weight." Take, again, the example from above: If an algorithm indicates a high risk for developing Parkinson's, but MRI scans, lab tests and the clinical history do not support this claim, a lack of interpretability makes it impossible to fully evaluate its outcome. Such an evaluation includes answering questions like: Are the various results really inconsistent or is there an integrative interpretation available? If not, is there any alternative reading that brings about a consistent explanation? Does the logical relation of the different results suggest dismissing one of them as most probably wrong?

Common intersubjective practice allows for various forms of gradation, repetition, revision, and distribution. We regularly doubt the validity of evidence provided to us by others and demand more or different substantiation. In such a case, our counterpart may simply repeat his initial claim, mitigate it or add further aspects and, thereby, densify the inferential web. While we acknowledge that the patient-physician relationship is intrinsically asymmetric many people place a high degree of trust in their physician, the critical aspect is that any patient can in any (nonemergency) situation ask for explanation and discourse. In fact, informed consent including the opportunity to ask questions is an essential part of the ethical and legal requirements for medical intervention. Nondiscursive elements are alien to this practice and it is impossible to integrate them neatly. The outcome of most state-of-the-art algorithms are monolithic in this sense. They do not have any connecting points that allow for gradations and revisions. They are "Take it or leave it"—things which are, ultimately, incompatible with our discursive practice if they should move beyond the status of current lab-tests, that is, information for the physicians that are weighted against other information.

It would be disproportionate to conclude from all this that we should not use machine-learning algorithms altogether and stop developing the tools for "medical AI." Presumably, such systems will become highly reliable for various tasks in the foreseeable future. It would be unreasonable not to use them, in particular, in the medical context for diagnosis and prediction. The challenge is, rather, to realize that they need to be connected to other information and find ways to make their outcomes compatible with our discursive practice. Ultimately, we need to find ways of integrating medical AI into our discursive practice. This will only be successful if medical AI is designed in a way that makes it integrable.

7 | EXPLAINABLE AI

Under the title of "explainable AI" initiatives heading in the direction just mentioned are already under way. Interestingly, the US-American Defense Advanced Research Projects Agency (DARPA) is among those promoting this development. In a report published on the DARPA website, David Gunning notices: "Dramatic success in machine learning has led to an explosion of new AI capabilities. Continued advances promise to produce autonomous systems that perceive, learn, decide, and act on their own. These systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to human users. [...] DARPA is interested in creating technology to make this new generation of AI systems explainable. Because the most critical and most opaque components are based on machine learning, XAI is focusing on the development of explainable machine learning techniques. By creating new machine learning methods to produce more explainable models and combining them with explanation techniques, XAI aims to help users understand, appropriately trust, and effectively manage the emerging generation of AI systems." (Gunning, 2016, p. 5–6) Moreover, big companies like Intel are paying attention to explainable AI. Casimir Wierzynski (Senior Director, Office of the CTO, Artificial Intelligence Products Group at Intel) observers: "Explainability is a scientifically fascinating and societally important topic that sits at the intersection of several areas of active research in machine learning and AI" (Wierzynski, n.d.). Finally, a symposium at the 2017 Conference and Workshop on Neural Information Processing Systems (NIPS) was dedicated to "Interpretable ML" (see http://interpretable.ml/; it is only by coincidence that there is a picture of Harry Potter and his friend Ron staring at a crystal ball on the front page of the symposium website). Nevertheless, the issue is only slowly being recognized as essential. In the *AI Now 2017 Report* it is hardly mentioned (Campolo, Sanfilippo, Whittaker, & Crawford, 2017). However, recently the EU's High-Level Expert Group on Artificial Intelligence has published a report on "Trustworthy AI" which includes an assessment list for evaluating AI systems (High-Level Expert Group on Artificial Intelligence, 2019). Under the heading "Transparency" explainability is listed as a key requirement (First et al., 2018, p. 29). In research on machine learning for medical diagnosis and prediction explainable AI does not currently play a significant role. As noted above, models detailing their reasoning in plain language are still unattainable. Explainability, however, can be achieved on the model level (a human's ability to understand the structure of the process, which provides a bridge to shared decision-making) and on the level of the results (why was this particular decision made in this specific case). This gap needs to be addressed before "medical AI" systems can be (or start to be) deployed in clinical practice. Most critically, machine-learning algorithms for medical purposes should be required to implement elements of explainable AI before getting approval by regulatory authorities, even if it will be difficult to specify a necessary threshold for such elements. Inferential vagrants are certainly not what we should like to see in medical practice. Physicians need to be able to inferentially authorize their diagnoses and predictions and take responsibility for them.

8 | IMPLICATIONS

For the moment, it is difficult to pin down what the requirements for machine-learning algorithms exactly entails—in view of integrability with other information, defensibility, and interpretability—and how it can be implemented best. Extensive research is needed for finding adequate solutions. From an ethical perspective compatibility with our discursive practice is the essential point. This is, at first, a conceptual claim. It demands that outcomes generated by an algorithm must include discursive elements or points of contact for linking them to other information and sources of knowledge und, by doing so, enable a weighting of information based on the presented evidence.

However, it is also a claim that calls for empirical investigation: What type of opacity in medical test results do people actually accept and which do they consider to be suspicious? Does it depend on the stakes, that is, the degree to which a diagnosis or treatment will affect an individual life? We may be more happy to accept a computer generated diagnosis that has only the consequence of needing to take a pill each morning than one that would require extensive surgery. This, too, is important in view of discursive practice. Therefore, studies including different target groups (patients, physicians, other medical staff) are urgently needed that address this question and the preferred scope of the explainability-requirement relative to severity, available therapies and other real-life consequences like becoming ineligible for insurance due to not yet present disorders. Comparing the reactions toward machine learning algorithms with those toward "expert intuitions"—which can, of course, also be epistemically opaque—might also be instructive. At any rate, profound knowledge of relevant factors will be essential for ethically acceptable approval procedures. In a nutshell, machine-learning algorithms call for deeper insights into the epistemology as well as the psychology of understanding.

9 | SPECIFIC CONSIDERATIONS ON NEUROIMAGING

Most of the difficulties and potential solutions for these discussed in this work are rather generic to medical AI. In the following, however, we will add a few thoughts that are directed more specifically to the field of neuroimaging, where machine learning based on mainly MRI but also EEG and PET data have received much attention over the past years (for detailed reviews see Arbabshirani et al., 2017, First et al., 2018, Wolfers, Buitelaar, Beckmann, Franke, & Marquand, 2015). Such approach has a particular appeal in the context of psychiatry where classical tests and objective biomarkers are largely lacking, and indeed many of the papers discussed in the previously mentioned reviews provide encouraging results. Compared to other application in medicine and other fields, though, sample sizes trend to be rather small in neuroimaging given the logistic expenses associated with scanning large cohorts of patients and healthy subjects. Worst yet, it has been noted, that prediction accuracy trends to decline with increasing sample size, which counters the logical expectation and could be an indication of overly optimistic procedures in earlier, smaller studies. This highlights the need for more independent testing of predictive algorithms, ideally using a new, unrelated dataset once all optimization has finished and the final model has been trained. At the moment, such approach is clearly limited by data availability, leading to a dominance of cross-validation over external evaluation (Mendelson, Zuluaga, Lorenzi, Hutton, & Ourselin, 2016; Varoquaux, 2018; Whelan & Garavan, 2014). We envision, though, that independent tests will become more of a norm in the future and that this development will have a positive mutual influence with increased data sharing—either fully open or through joint publications as is already the standard in genetics. Likely, accuracies will turn out to be lower in this case, but more reflective of the true level of generalization. That is, the true predictive power and hence most likely the limitations of machine-learning approaches for prediction in clinical neuroimaging will most likely only be revealed once the field is moving toward a more rigorous, out-of-site testing performed after completion of all model selection, hyper-parameter optimization and model training has been completed.

When further considering the fact that accuracy for diagnostic decisions (patient vs. control) are rarely exceeding 80–85%, apart from the case of Alzheimer's disease, it feels as if fully automated decision-making in brain medicine is unlikely to happen in the near future. This situation, however, poses the question of how to develop brain measures into clinically useful biomarkers for shared decision-making. That is, assuming that an autonomous diagnostic or prognostic assessment is not currently foreseeable, the information that can undisputedly be gained from neuroimaging data through machine-learning approaches should find another avenue into clinical practice. A conceivable role for predictive modeling would this to provide a score-sheet similar to a lab chart or the report from a diagnostic procedure. Such sheet could provide classification labels and associated likelihoods as well as reference information on the diagnostic accuracy of the algorithm in well-conducted (see above) evaluation studies for multiple disorders. For example, a patient with suspected Parkinson's disease could provide the labels, scores or probabilities for Parkinson's, Depression, Alzheimer's and other disorders as derived from the patient's MRI scan. The consulting physician could then integrate and weight this information relative to the clinical history, their own examination, lab results, and other evidence. We would argue that this would just shift the need for transparency and interpretability from the patient to the physician, who will most likely put greater importance on the algorithmic findings if they can be understood and hence put into discourse with the surrounding information.

This also leads to a related yet rarely explored field, namely the integration of machine learning with existing knowledge about human brain organization and (nonimaging) pathophysiological understanding. Such integration will most likely be bi-directional, as on the one hand prior knowledge can be powerful for feature definition and reduction of the search space whereas on the other relevance of a given feature of predictive power of a certain set of features can provide new insights into the organization of the brain and its pathology (Chin, You, Meng, Zhou, & Sim, 2018; Nostro et al., 2018; Weis et al., 2019). Similar considerations may hold for the aforementioned juxtaposition of clinical examination, test reports and machine-learning outputs, which should lead to novel insight and hopefully a consequential refinement of machine-learning algorithms based thereon. That is, predictive algorithms can both benefit from prior information and feedback relevant insight about diagnostic relevance of a given feature set. In this context, we would clearly position ourselves against the idea that "the data will sort itself out" given that it seems unrealistic that any learning approach will ever see enough data to adequately cover the long and fat tails of clinical distributions.

A final question for machine learning on MR imaging data arises from the broad range of different imaging sequences and hence features that can be gained. Currently, most work on is geared toward the use of either structural imaging data or resting-state fMRI for classification. Yet, even within either of these, many different features can be extracted, for example, local volume, shape and cortical thickness from structural MRI, functional connectomes, ICA maps or local measures such as regional homogeneity (ReHo) or the amplitude of low frequency fluctuations (ALFF) from resting-state data. In addition, diffusion-weighted imaging, task-based fMRI including naturalistic stimulation and arterial spin labeling (ASL) all provide suitable features. Given that each of these reflect a different, distinct part of individual neurobiology and presumably pathophysiology, it seems reasonable to assume that the most relevant features may not be likewise present in all modalities. From a theoretical perspective with unlimited training data, multimodal imaging features would thus allow researchers to always find classifiers that are at least as good as those from any individual modality. If relevant features would only be present in one modality, it would pick them in the same way as if only this modality was available and yield the same accuracy. If they were distributed across modalities, it could combine them to beat any unimodal one. Several approaches for multimodal classification have thus been employed on neuroimaging data, including concatenation, multi-kernel learning or, as a more recent development through the synergy rule, which is constructed based on an integration of multiple related monotonic, for example, SVM, classifiers (see for examples Liem et al., 2017; Schmaal et al., 2015; Vapnik & Izmailov, 2016; Youssofzadeh, McGuinness, Maguire, & Wong-Lin, 2017). It remains to be seen, though, whether machine learning based on multimodal MR imaging data can outperform unimodal ones in strictly independent evaluations, given the higher number of features and (usually for logistic reasons) lower number of observations for the former, which may lead to over-optimistic assessments as discussed above.

In summary, potentially owing to the absence of objective tests for many brain disorders, neuroimaging has been very keen to explore the use of machine-learning tools with the aim of clinical applications. While yet limited by low sample sizes and potentially over-optimistic accuracies due to a lack of external validation sets, these approaches may become valuable assets in a shared decision-making context that may at the same time build upon and feed into knowledge on brain organization and pathophysiology.

10 | ADDENDUM

In the scene from "Harry Potter and the HalfBlood Prince" mentioned at the outset Harry shies away from disclosing the evidence he has for his accusation of Malfoy because of his deep mistrust of Professor Snape. Deliberately not providing evidence can occasionally be an act of taking responsibility—but one that is open only to discursive beings who can, in principle, give and ask for reasons.

CONFLICT OF INTERESTS

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Bert Heinrichs https://orcid.org/0000-0002-0181-0078

REFERENCES

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Bondy, P. (2015). Epistemic Value. *The Internet Encyclopedia of Philosophy*. https://www.iep.utm.edu/ep-value/

Brandom, R. B. (2000). *Articulating reasons: An introduction to Inferentialism*. Cambridge: Harvard University Press.

Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in Neurosciences*, *42*(4), 251–262. https://doi.org/10.1016/j.tins.2019.02.001

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry*, *3*, 223–230.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017): *AI Now 2017 Report*. https://ainowinstitute.org/AI_Now_2017_Report.pdf

Castelvecchi, D. (2016). Can we open the black box of AI. *Nature*, *538*, 20–23.

Chin, R., You, A. X., Meng, F., Zhou, J., & Sim, K. (2018). Recognition of schizophrenia with regularized support vector machine and sequential region of interest selection using structural magnetic resonance imaging. *Scientific Reports*, *8*(1), 13858. https://doi.org/10.1038/s41598-018-32290-9

Choudhary, R., & Gianey, H. K. (2017). Comprehensive review on supervised machine learning algorithms. *Proceedings of the First International Conference on Data Science, E-learning and Information Systems (MLDS)*, Noida, India, *2017*: 37–43.

Council Regulation. (2016). Regulations. *Official Journal of the European Communities* No L 119/1, 4.5, *2016/679*, 1–88.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115–118.

Eun-Jae, L., Yong-Hwan, K., Namkug, K., & Dong-Wha, K. (2017). Deep into the brain: Artificial intelligence in stroke imaging. *Journal of Stroke*, *19*, 277–285.

First, M. B., Drevets, W. C., Carter, C., Dickstein, D. P., Kasoff, L., Kim, K. L., … Zubieta, J. K. (2018). Clinical applications of neuroimaging in psychiatric disorders. *The American Journal of Psychiatry*, *175*(9), 915–916.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, *23*, 121–123.

Gunning, D. (2016). *Explainable Artificial Intelligence (XAI)*. DARPA-BAA-16-53. Broad Agency Announcement. Arlington: Defense Advanced Research Projects Agency, Information Innovation Office. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2018). *Do deep neural networks* *outperform kernel regression for functional connectivity prediction of behavior?* https://doi.org/10.1101/473603

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Brussels.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*, 428–434. https://doi.org/10.1016/j.tics.2007.09.004

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*, 615–626.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Janssen, R. J., Mourao-Miranda, J., & Schnack, H. G. (2018). Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*, 798–808.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*, 255–260.

Kelp, C. (2015). Understanding phenomena. *Synthese*, *192*, 3799–3816.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspectives. *Artificial Intelligence in Medicine*, *23*, 89–109.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge: Cambridge University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., … Margulies, D. S. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, *148*, 179–188.

Maier, A., Schebesch, F., Syben, C., Würfl, T., Steidl, S., Choi, J., & Fahrig, R. (2018). Precision Learning: Towards Use of Known Operators in Neural Networks. *24th International Conference on Pattern Recognition (ICPR)*, Beijing. *2018*: 183–188.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mendelson, A. F., Zuluaga, M. A., Lorenzi, M., Hutton, B. F., & Ourselin, S. (2016). Alzheimer's Disease Neuroimaging Initiative. Selection bias in the reported performances of AD classification pipelines. *Neuroimage: Clinical*, *14*, 400–416.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., & Alzheimer's Disease Neuroimaging Initiative. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412.

Nostro, A. D., Müller, V. I., Varikuti, D. P., Pläschke, R. N., Hoffstaedter, F., Langner, R., … Eickhoff, S. B. (2018). Predicting personality from network-based resting-state functional connectivity. *Brain Structure & Function*, *223*(6), 2699–2719.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Proceedings of the First International Conference on Data Science, *E-learning and Information Systems (MLDS), Noida, India*, 2017

Quine, W. V., & Ullian, J. S. (1978). *The web of belief* (2nd ed.). New York: McGraw-Hill Education.

Schmaal, L., Marquand, A. F., Rhebergen, D., van Tol, M. J., Ruhé, H. G., van der Wee, N. J., … Penninx, B. W. (2015). Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: A multivariate pattern recognition study. *Biological Psychiatry*, *78*(4), 278–286. https://doi.org/10.1016/j.biopsych.2014.11.018

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*, 1085–1139.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*, 233–242.

Vapnik, V., & Izmailov, R. (2016). Synergy of monotonic rules. *The Journal of Machine Learning Research*, *17*, 4722–4754.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*(Pt A), 68–77.

Wachter, S., Mittelstedt, B., & Floridi, L. (2017). Why a right to rxplanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*, 76–99.

Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., & Eickhoff, S. B. (2019). Sex classification by resting state brain connectivity. *Cerebral Cortex*, bhz129. https://doi.org/10.1093/cercor/bhz129

Whelan, R., & Garavan, H. (2014). When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biological Psychiatry*, *75*(9), 746–748. https://doi.org/10.1016/j.biopsych.2013.05.014

Wierzynski, C. *The Challenges and Opportunities of Explainable AI*. https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/.

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews*, *57*, 328–349. https://doi.org/10.1016/j.neubiorev.2015.08.001

Youssofzadeh, V., McGuinness, B., Maguire, L. P., & Wong-Lin, K. (2017). Multi-kernel learning with Dartel improves combined MRI-PET classification of Alzheimer's disease in AIBL data: Group and individual analyses. *Frontiers in Human Neuroscience*, *11*, 380. https://doi.org/10.3389/fnhum.2017.00380

How to cite this article: Heinrichs B, Eickhoff SB. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum Brain Mapp*. 2020;41:1435–1444. https://doi.org/10.1002/hbm.24886