Ethics and Information Technology (2022) 24:22
https://doi.org/10.1007/s10676-022-09629-y

ORIGINAL PAPER

Epistemo-ethical constraints on AI-human decision making for diagnostic purposes

Dina Babushkina[1] · Athanasios Votsis[2]

Accepted: 5 January 2022 / Published online: 19 April 2022

Abstract

This paper approaches the interaction of a health professional with an AI system for diagnostic purposes as a hybrid decision making process and conceptualizes epistemo-ethical constraints on this process. We argue for the importance of the understanding of the underlying machine epistemology in order to raise awareness of and facilitate realistic expectations from AI as a decision support system, both among healthcare professionals and the potential benefiters (patients). Understanding the epistemic abilities and limitations of such systems is essential if we are to integrate AI into the decision making processes in a way that takes into account its applicability boundaries. This will help to mitigate potential harm due to misjudgments and, as a result, to raise the trust—understood here as a belief in reliability of—in the AI system. We aim at a minimal requirement for AI meta-explanation which should distinguish machine epistemic processes from similar processes in human epistemology in order to avoid confusion and error in judgment and application. An informed approach to the integration of AI systems into the decision making for diagnostic purposes is crucial given its high impact on health and well-being of patients.

Keywords Hybrid epistemology · Ethics and epistemology of AI · Fuzzy concepts · Medical AI · AI in decision making

First take on the problem: AI, moral decision making, and uncertainty

Assume you do not know anything about Artificial Intelligence (AI) except what is out there in the air. Assume you do not understand how neural networks work.[1] You may have seen various colorful diagrams of nodes and layers connected by arrows. In deep learning methods as applied to pattern recognition, these diagrams with "neurons" and weights are learning some incomprehensible magic on images, compressing, multiplying, going back and forth, and then in the end you have a result, say, "85% dog, 15% wolf". You may have come across various explanations, you may even be able to draw an answer if someone asks you what this algorithm is doing, in general terms. But that is all you know. Most of us, including direct users (people using AI systems in devices in their everyday lives) and professional users (who may rely on AI technology for performing their professional tasks, e.g. doctors),[2] are, in the best case, at this level of understanding. Moreover, the training of that portion of users who do have experience with quantitative methods is arguably heavily or solely stemming from frequentist statistics.

Now let's assume that based on this result *alone*, you have to make a decision on whether you will take an animal into your care. Can you make this decision based on the outcome of the algorithm? Or rather, would it be a

✉ Dina Babushkina
d.babushkina@utwente.nl

[1] Section of Philosophy, Faculty of Behavioral, Management and Social Sciences, University of Twente, Enschede, The Netherlands

[2] Section of Governance and Technology for Sustainability, Faculty of Behavioral, Management and Social Sciences, University of Twente, Enschede, The Netherlands

[1] In this paper we focus on deep neural networks (DNN) because of their pervasive use for medical diagnosis, as well as health app technologies that are increasingly introduced into the extra-clinical domain. In principle, we anticipate that our argument, with appropriate alterations, will stand in the cases of other types of neural network architectures (e.g. shallow/sparse neural networks). However, we advise that different types of machine learning must be independently analyzed in terms of their epistemology and role in the hybrid decision making process for diagnostic purposes.

[2] On the rise of AI in healthcare applications see e.g. Bohr and Memarzadeh (2020). For AI/ML-based medical systems approved by the US Food & Drugs Administration, see Benjamens et al. (2020). Topol (2019) is one of the most extensive arguments for the use of AI in healthcare.

good, well-grounded decision? Let's assume, for clarity's sake, that you want to make a present for your kid—a dog that he/she has been asking for. Now, the question here is not whether it is possible for you to make such a decision. People make all sorts of decisions, not all of them are rational and well-grounded. The question here rather is: *does the output of the algorithm (This animal is 85% dog and 15% wolf) constitute a good justification for your decision to bring the animal home to your child, and given the uncertainty, whether it is rational for you to take the risk that the animal might turn out to be a wolf?* Turn it to be a wolf, you will have to deal with the consequences of having a wild animal in your house. This not only means frustrated expectations of your child and a considerable impact on your life, but most importantly this will increase the possibility that someone in your home will come to harm and that the animal will suffer. In the end, you will have to weigh your trust in the algorithm against the ways in which the uncertainty affects the alternative futures, including your responsibilities, other people involved, and the destiny of the animal. This example may not appear so dramatic: there are a number of ways you could have avoided dealing with an unfavorable outcome. But imagine that the consequences of your choice would be far more severe than just adopting an unwanted animal. Say, that based on the result the animal will be killed or let live. How should the AI's 85/15 ratio weigh in your decision process now? What if that was not an animal, but a human being? Imagine now that you are a medical doctor and you are looking at the results of an AI system that tells you that the patient in your care has 85% pneumonia, 10% cancer, and 5% tuberculosis (further: 85/10/5). How should you weigh this ratio against the consequences of your decision concerning the patient's diagnosis for his/her health, well-being, and potentially life?

The problem we are dealing with stems from (a) the fact that the outcome of AI's calculation figures *as a reason* in the decision making process, and (b) the fact that it is an open question whether *this reason is any good*. To ask this question is especially important, given the weight of uncertainty due to the epistemic contribution of an AI system against the possible harm. What is at stake here is not so much the question about what considerations it is sensible for the decision maker to count as reasons for action, given that he/she cannot be certain about some key parameter, P, relevant to the case. That is to say, not so much whether, given that you do not know if the animal in question is a dog or wolf (similarly, given that a medical professional does not know for sure what condition the patient is suffering from), it is sensible for you (the doctor) to align your (his/her) decision with the output of the AI system. This may be the case, under some circumstances which we do not need to discuss here. The question is rather what matters *ethically* from the perspective of one potentially harmed by the decision based on such considerations: for your child—whether the animal is indeed a dog and thus having it in the house won't be associated with danger; for the patient—whether the condition she is suffering from is indeed pneumonia, and that she will receive fitting treatment. Of direct ethical relevance here are the consequences of the mistake that results from using AI output as a reason for taking a certain course of action towards the patient, such that would affect her health and well-being (e.g. wrong treatment, extra suffering, possible death).[3] Seeing the role of AI in decision making from this perspective helps answer the questions: How should AI results be interpreted during decision-making, when such a decision entails risk of harm and significant moral cost? How to minimize the risk of misdiagnosis when the diagnostic procedure is aided by an AI tool?

One aspect of this is minimizing the risk of harm due to a wrong decision and making sure that you do not just guess. That entails making sure that if a consideration counts, for you, as a reason pro or against a certain decision, *it is suited to be such a reason*. Normally that presupposes making an effort to find out whether its validity can be independently verified. The question here should be: is my expectation that this consideration constitutes a good reason for action itself reasonable? Is my anticipation, that the consideration represents an expert opinion or that this consideration is a result of certain capacities or skills, backed up by reality or is it a mere assumption?

Another important aspect is understanding what sort of uncertainty is figuring in the particular situation where you are making your decision, which methods are appropriate to address this type of uncertainty and, as a result, which considerations can count as good reasons. If we take the role of an AI decision support system to be the minimization of the probabilistic uncertainty of concepts as "event occurrences", that would entail certain constraints on the machine inference process. The decision here is supported by quantified information about the uncertainty of whether an event is likely to occur or not, given information on past occurrences and explanatory and control factors. This is firmly rooted in a set of epistemological assumptions of clearly delineated, mutually exclusive, and repeating events. There is, however, a key departure from these in medical decisions. Strictly speaking, what

[3] Here we focus on one aspect of the application of AI in healthcare. The spectrum of possible ethical problems that are associated with AI in healthcare is well described in research literature. We refer readers, e.g., to Morley et al. (2020).

is uncertain is not whether an event occurs or not, but whether the decision maker (e.g. a medical professional) is confronted with a case that fits the profile of disease A versus disease B. This has resemblance to the identification problem[4] in probabilistic reasoning but here it transcends the domain of model specification; instead of misidentified effects of a variable on a phenomenon, the issue is the kind of phenomenon that the model treats altogether. Furthermore, there is uncertainty about the ontology of A and B: their features may overlap, it is not clear how a pool of explanatory factors will lead causally to A as opposed to B,[5] and in some cases frequency of occurrence cannot be established as the 'events' appear once without prior information because they constitute one-off unique cases. This sort of epistemological uncertainty, in the domain of diagnosis, moves models away from the task of inferring the probability of a phenomenon occurring to that of whether it is possible—or whether one should believe—that the logical inference of the model is this or that phenomenon altogether.

Second take on the problem: hybrid epistemology

Cases like AI assisted diagnosis are no longer bound *merely* by the norms of human epistemology. Here we are dealing with a new phenomenon of *hybrid epistemology* (or "hybrid intelligence" as in van Baalen et al., 2021). By hybrid epistemology, in this context, we understand an intimate mix of human cognitive processes and machine procedures that manipulate and transform information in uniquely different ways. One pitfall to avoid when assessing these procedures and the role they should play in decision making, is getting trapped in the predominant narrative about AI and continue extrapolating terminology describing human cognition. The existing research is still over-reliant on the aspirational narrative of replicating human cognition in an artificial environment (*de dicto*). As a result, the narrative about AI is still predominantly metaphorical, reflecting what we would like algorithms to do. Any claim about the role of AI in decision making that rests on the analogy between human and machine is bound to beg the question. One first has to prove that the analogy stands and explain to what extent. We need to realistically assess machine epistemology and construct a conceptual apparatus that would do justice to the unique elements of such epistemology, reflecting what algorithms in fact do (*de re*). There is a pressing need to analyze the epistemic capabilities of different types of algorithmic solutions, estimate how these capacities relate to the production of knowledge, and deduce normative constraints that apply to them. This is crucial if we are to realistically assess what sort of conclusions we are warranted to draw from AI algorithms. On the positive side, the urgency of such research has become apparent. One of the overarching aims of this paper is to problematize hybrid epistemology and contribute to the demarcation of the specific sub-tasks in such a system, with clear understanding of the epistemic vulnerabilities of the human and machine components as well as the inference from the latter to the former. This should help to draw the line between the specific epistemic responsibilities of each component (on the need to clearly delineate the epistemic tasks of AI and the human expert in the hybrid intelligence cf. van Baalen et al., 2021; Boon, 2020), and to minimize "trade-offs at the epistemic and the normative level" (Grote & Berens, 2020).

With respect to the problem of hybridization of decision making, one of our central claims is that AI systems are not substituting an element of human epistemology. AI is not imitating a human cognitive faculty but is creating a unique epistemological product.[6] So, the main question here is, given the asymmetry between AI and human epistemology, how are we to *appropriately integrate*[7] this product in the human decision making process? What are the

[4] The problem of not capturing and measuring the effects of the intended concept or attribute through a certain variable, because the variable reflects something different or a mixture of different things that what the researcher assumed. Any inference will be therefore problematic due to misidentification of effects.

[5] On the causal roots of probability see Belis (2007).

[6] An anonymous reviewer has drawn our attention to a potential objection: given that science still does not know the mechanisms that underly human cognitive mechanisms (such as drawing conclusions), it is possible that human cognitive processes are comparable to those of DNN and subject to the same criticism as this paper puts forward against the machine inferences. This is an interesting but controversial topic, which requires a thorough investigation of debates in philosophy of science and philosophy of mind. We cannot attempt to explore these in detail here. However, to clarify our position in this paper, we take a normative approach to the question of cognition, drawing from epistemology as a philosophical discipline, which, simply put, studies rational (i.e. universal and a priori) constraints in the concept of knowledge and its forms. The question about mechanisms in the brain that underly human cognitive processes (such as thinking, imagining, or perceiving) is a descriptive question. The information about the latter does not necessarily affect the truth about the former. However, if it is true that we do not know exactly how the human brain generates knowledge (descriptive domain), there is a very slim chance that calculation processes of the current DNN are, in fact, exactly the same with the processes in the human brain, unless this has happened *by chance*. In any case, in a situation like this, the burden of proof is on the side of the objector, i.e. one who claims that these two cognitive systems are, in fact, identical.

[7] Despite being framed in different terms, augmented rather than hybrid decision making, Jussupow et al. (2021) offer interesting insights about meta-cognitive skills required for the incorporation of AI output in the diagnostic process of a medical professional.

epistemological constraints on such integration? The importance of this discussion is especially obvious in cases like the use of AI tools to aid diagnosis, care and treatment, given the high personal cost that mistakes in these domains have for the patient. Another aspect of this is the fact that we cannot simply assume that there is a symmetry between an AI's output and a certain type of human judgement. Any such assumption has to be checked against the facts concerning the processes involved in a specific type of an algorithm, and without a sufficient reason to establish an analogy between any machine process (say, a deep neural network outputting a certain ratio) and a human cognitive process (say, a judgement about the probability of a certain event), no such assumption should be accepted.

Hybrid epistemology for decision making purposes in health care: an example

In general terms, the problem with hybrid decision making that we have been introducing thus far can be schematically presented as follows (Fig. 1)[8]:

AI MODEL OUTPUT → HUMAN INFERENCE → HUMAN DECISION → PRACTICAL CONSEQUENCES[8]

Fig. 1 Schematic representation of hybrid decision making

An example of hybrid epistemology for decision making purposes in health care could look something like this:

1. An output of an AI system, designed to assist diagnosis of lung condition.
2. A judgment by a human medical professional, based on the output of the AI system, as to the condition of the patient.
3. A decision about the appropriate treatment of the patient.

A process similar to this would underlie the decision procedure of medical professionals involved in the application of systems like CheXpert Analysis (Irvin et al., 2019) or CheXNet (Rajpurkar et al., 2017). Both systems process X-ray images. In the former case the medical specialist sees output of the type "pleural effusion (very likely), abnormal (very likely), atelectasis (likely), cardiomegaly (unlikely)". In the case of the latter, the medical specialist sees a binary result, e.g. "pneumonia positive (85%)".[9] Application areas of AI in healthcare are diverse and include ophthalmology (e.g. screening for glaucoma, diabetic retinopathy, hypertensive retinopathy; cf. Gulshan et al., 2016; Poplin et al., 2018), dermatology (e.g. skin cancer, cosmetic care; cf. e.g. Elder et al., 2021; Rundle et al., 2021), mammography (see e.g. Mayo et al., 2019; Wu et al. 2019; Badré et al. 2021; Liu et al. 2021; Sheth & Giger, 2020), and pulmonology (a range of acute and chronic conditions; see e.g. Kaplan et al. 2021; Almaslukh, 2021; Almalki et al. 2021). Such research and development has been also finding its way to online education, for instance via entire specializations for AI in the medical sphere. For instance, on Coursera, the specialization by DeepLearning.AI includes the application of such models in the domains of diagnosis, prognosis and treatment, and does not require prior medical training. This gives a good idea about how the AI assistant could function and what sort of input the AI model will offer to the decision making process concerning diagnosis.

The type of AI system that we are talking about in the context of this paper is based on deep neural networks (DNN) as applied to image recognition (visual patterns). For simplicity's sake, we will use a hypothetical case, similar to the one generated by the system in Irvin et al. (2019). In our example, a medical professional receives the following result from his/her AI diagnostic assistant: "85% pneumonia, 10% cancer, and 5% tuberculosis". An intuitively appealing mode of hybrid reasoning in this case would be something like this:

> [α]: [α1] if an output of an AI algorithm is ("85% pneumonia, 10% cancer, and 5% tuberculosis"), then it is *most likely* pneumonia,

[8] This is, of course, a purposeful simplification of the hybrid decision making process. The goal here is to isolate and analyze the elements which are directly relevant to the problem discussed in the paper. We chose to focus on this element of the decision process because it illuminates the specific danger at hand, i.e. overreliance on AI diagnostic tools. What we think is especially important at this stage—when an explosion of new AI-based diagnostic software happens but the understanding of their limitations and proper ways to integrate them into the wider diagnostic process are lagging behind—is to prevent the situation where AI output is the only thing that is taken into account when making a diagnosis or where this output is by default given more weight. This motivated us to isolate this specific part of the hybrid decision making process and show (a) that there is a high danger of incorrectly interpreting the nature of AI output, given its deceptive or superficial representation in specialized software, and (b) that such output cannot be taken outside the context. Our goal is to show that AI diagnostic software must always be evaluated by a medical professional(s) in light of additional evidence. We are grateful to an anonymous reviewer for drawing our attention to the need to add this clarification

[9] For an overview of AI in medical diagnostic imaging see, e.g. Kapoor et al. (2019), Fujita (2020), Ting et al. (2021).

> [α2] therefore [for the diagnostic purpose] it is pneumonia.

The first part of this reasoning process [α1] represents *some form of probabilistic inference*. One premise in this inference is the output of the AI-based diagnostic assistant system (it can be seen as an AI-generated knowledge content, i.e. that, which later will become the content of human knowledge); the second part [α2] is an inference about a matter of fact. This probabilistic inference plays a crucial role in the decision making process: it figures as a reason that weighs towards a certain course of action. This reason sanctions to act *as if* the patient indeed was affected by pneumonia. This dialectic between complex epistemology and ontology that has concrete practical consequences needs further investigation. Before we can accept it as a reason justifying a course of action, we have to establish whether [α2] is warranted, by what, and to what extent. In other words, the question is: *given the ethical constraints of the decision making situation at hand, to what extent are we justified to move from* [α1] *to* [α2]?

It is helpful to break down the reasoning process [α] into more specific steps:

> [claim]: The output of the AI diagnostic assistant tool is "85% pneumonia, 10% cancer, 5% tuberculosis".
> [inference 1]: Therefore, the patient is *most likely* suffering from pneumonia.
> [inference 2]: Therefore, it is a case of pneumonia that we are dealing with.
> [conclusion]: The patient should be prescribed a treatment for pneumonia.

The initial claim isolates the element of machine epistemology, while the rest of the steps represent human reasoning. Now, our conclusion is warranted only when the intermediate steps (i.e. the moves from inference 1 to inference 2 and, in turn, from the initial claim to inference 1) are warranted. To understand whether we can move from 85/10/5 to "most likely x", we have to find out what 85/10/5 actually means. This, in turn, is determined by the epistemological process by the means of which the algorithm arrives at this ratio. This is what we will be engaged in for the rest of the article: in the first part we will show how one should not interpret the AI output. For this, we will reconstruct two intuitively appealing and most probable interpretations and then discredit them in the light of the knowledge about the processes by which the AI system arrives at its output. In the second part of the article, we will make our suggestion about a more plausible way to interpret the AI's output and fitting ways to incorporate it into the decision making process.

How not to interpret AI output

"Most likely"

If we take a closer look at inference 1, it is reasonable to assume that this "*most likely*" may be interpreted in one of two alternatives:

(1) A claim about a phenomenon *fitting a certain concept*. It would read something like this: "This case is very much like pneumonia, and not so much like cancer or tuberculosis".
(2) A claim about *an event likely to occur*. "85% x" would translate into something like this: "In 85 out of 100 cases x was the signature of pneumonia".

Despite their intuitive appeal, both interpretations, as we will show, are erroneous. But first we need to say a few words about why both are so intuitively appealing. In a sense, they both are an expected consequence of the ambiguous language of the standard machine (or deep) learning narrative. This narrative is not only largely metaphorical (i.e. not literary applicable) when it comes to the attribution of human cognitive skills to an artificial system, but—more importantly in this context—is not consistent in maintaining the analogies with human cognitive skills. Let us explain.

The probabilistic inference of the type "AI output is 85% x, 10% y, 5% z, therefore it is most likely x" is based on a naïve assumption about the processes leading to an AI system's output. It is this naïve assumption that the illusion of the plausibility of this inference relies on. The naïve assumption that appears to support this inference goes something like this:

[β] The AI system learns to identify x from examples of x to which it is exposed, and then uses the derived knowledge of x to identify a new instance of x.

This naïve interpretation is encouraged by the meta-description of DNN, namely, of the "training" and "prediction" stages that DNN is said to be split into. The spell of analogy compels the standard AI narrative to tag the stages of the algorithm akin to the stages of a human learning agent. The training stage is explained as the part of the algorithm where machine learning happens: the DNN is said to extract the features representative of a certain phenomenon and fix it in a mathematical formula. The prediction stage is presented as the part of the model which applies the pattern it identified to previously unknown samples. During this stage, the AI is said to process previously unknown cases and identify it as belonging (or not) to the same class. This standard AI terminology is stretched to mimic basic human cognitive processes necessary for the identification of a new instance of something:

[γ] observation of existing cases → abstraction of relevant features → application to new cases

In a very rough explanation, the human process involves learning, which is based on induction (involving abstraction, generalization, and the extrapolation of a principle) and the subsequent deduction (i.e. an application of the general principle to a new instance). The problem is that in human cognition this process can take one of the two basic forms:

[γ1] observation of existing cases → concept → application to new cases

[γ2] observation of earlier occurrences of the same phenomenon → rule → prediction of new occurrences

The former is suitable for an a-temporal identification of entities (things), while the latter is about events occurring in time. If one wishes to follow through the analogy with the human cognition, the challenge is to interpret the DNN stages and underlying processes in terms of [γ1] and [γ2]. This is, however, a nearly impossible task, since the meta-narrative about AI techniques does not seem to distinguish between [γ1] and [γ2], effectively blending these two cognitive procedures. The narrative about "the training stage" appears to suggest that the algorithm is forming some sort of concept (in our first example, of a "dog" and a "wolf" and in our second example, of "pneumonia", "cancer" and "tuberculosis", since the algorithm is said to be trained to recognise instances of each of these). However, "the prediction stage", as the very name suggests, is supposed to be about the prediction of an event. This, of course, opens the way for confusion, inviting erroneous interpretations and false expectations. It is this blending of two distinct human cognitive processes that explains the appeal of the intuitive interpretations of "most likely" (i.e. the concept fitting and the event prediction).

In the light of this, it is important to evaluate whether either of these interpretations are justified. And the way to do so is to clarify how AI arrives at the output it does. The machine arriving to its output must be properly positioned in the general machinery of hybrid decision making, which now looks like this (Fig. 2):

DATA PROCESSING → AI MODEL OUTPUT → HUMAN INFERENCE → HUMAN DECISION → PRACTICAL CONSEQUENCES

Fig. 2 Amended overview of the hybrid decision making process

In what follows, we will focus on the first node of Fig. 2, in order to show that neither of these interpretations, despite their intuitive appeal, are justified. We will explain how the algorithm works,[10] without falling into the trap of metaphors (*de re*, and not *de dicto*).[11] This will help to construct an account that fits machine epistemology. First, we will introduce and explain what a *machine concept* is (distinguished from a concept as an element of human epistemology), and explain the process involved in the construction and the subsequent application of any machine concept. Next, we will argue that in this type of hybrid epistemology, AI models are assigned an epistemological role that diverges prohibitively from how *the predictive output* of those same AI models is interpreted while inference and decision is being made. In somewhat simplifying terms, AI models are given the capacity to generate information about one type of uncertainty, but in the end are asked to support decisions by interpreting them from the standpoint of another type of uncertainty.

Machine concepts and the signification problem

The naïve interpretation [β] of what a DNN model does is based on at least two basic assumptions:

> [β1] that the AI system finds a way to identify *something*, i.e. an entity or an event. That is to say, that the output of an AI system *is about a certain phenomenon*, such as dog (or pneumonia), and
>
> [β2] that the entity (or event) in question is such that it *can be identified, and identified reliably, by the means that the AI system uses*. This means two things: that a dog as opposed to a wolf or a cat (or that the case of pneumonia as opposed to a case of, say, cancer) can be singled out by such a method; and that when applying the method, one would be able to differentiate between animals (or lung conditions) in the future, i.e. were one to meet an unknown animal (or lung condition), the method would be sufficient to say with reasonable certainty whether it is a dog (or, in the alternative example, a case of pneumonia).

But does DNN really identify anything in this way? To answer this question, we need to understand what sort of interpretation of "85% pneumonia" (as an element in the broader statement "85% pneumonia, 10% cancer, and 5% tuberculosis") is warranted by the nature of the internal process of the algorithm, at each of its stages. This is what we will do now. We will refer to expressions such as "85% pneumonia" as *machine concepts*. In this context, we take them to be elementary building blocks of machine epistemology, which are, as we intend to argue, distinct form concepts as building blocks of human epistemology.

The assumption that the output of an AI system is about *a phenomenon* (such as pneumonia), once more, has its

[10] Due to the lack of space, we will only focus on the relevant elements of DNN. As a result, the description of the algorithm will necessarily omit some details.

[11] For a deeper look into the mechanism of DNN we refer the reader to e.g. Chollet (2018) (from an applied computer science perspective) and Sullivan (2019) (from an epistemological perspective).

roots in reasoning by analogy with human cognition. In that sphere, a word only then is said to mean anything, when it refers to something that is real, i.e. to a certain entity, and expresses an idea which serves as a mental representation of this entity. It is only because lungs can get affected by bacteria or a virus in a certain way (the signified: pneumonia as an entity E) and because people have formed an idea that reflects the specificity of this condition (the meaning), that the word "pneumonia" has the meaning that it has (the signifier). This is what is commonly referred to as the signification triangle (Ogden & Richards, 1923; Peirce, 1998). So, the reference to a certain entity, in this paradigm, would be necessary to guarantee that 85% means anything at all.

In the attempt to match the narrative of the human mind, the standard AI terminology creates an impression that the algorithm is producing something similar to the meaning of "pneumonia" and that, akin to human mind, this happens in the interaction between a concept (such as given by a human, a developer, and then confirmed by the expert opinion of a group of medical experts) and the phenomenon (confirmed cases of lungs affected by pneumonia). However, the machine semiotics play out in a very different setting. The signification of any machine concept like "85% pneumonia" happens between the label, data-carrier and pattern.

The first element of the signification involved in the production of meaning when it comes to machine concepts like "85% pneumonia" is what we will refer to as a data-carrier (DC). The *data-carrier* is an easily understandable by a human digital representation of a certain real-life phenomenon. It can be a digital image (in our example an X-ray image of lungs), video, or audio file. An image is a data-carrier because its sole goal is to be a medium of data for the algorithm. *Data* is a mathematical representation of the information contained in the carrier. It quantifies the partial information about the entity. The way an AI system (in our case, DNN) comes into epistemic relation with the real life phenomenon is radically different from the way a human agent does: its relation to the object is always mediated by the process of digitalization, which serves the goal of translating reality into the symbolic language that a machine is capable of processing (codification). Digitalization is a procedure by the means of which the external world, as it is presented to humans, is made processable to an algorithm. There are a few important elements that are involved in the process of digitalization:

- Digitalization involves *fragmentation* of reality, that is to say, it effectively splits an entity (such as a dog or person) into spatio-temporal slices.
- Digitalization is a *reductionist act* because it reduces an entity to its fragments, i.e. to a limited selection of spatio-temporal slices or even one such slice (e.g. in our case, to an X-ray). As a result, it is necessarily selective and *discriminative* of other, potentially significant elements.
- However, at another level, digitalization is completely *indiscriminative* because it does not differentiate the entity from its environment. In machine epistemology, an image (recognized by a human as a picture of a dog) is a *totality T* (in which a human will distinguish between, say, sky, grass, leash and the dog itself). The epistemic noise that a human agent will have no problem discarding will figure in machine calculations on equal grounds with other features of the image.

A data-carrier is never the same thing as the entity which it represents. To begin with, it is different ontologically: the photo of the dog is not a dog. The same way, an X-ray image of the lungs affected by pneumonia is not the same thing as the actual sick lung. If not for any other reason, then at the very least, because the latter evolves and exists in the context of the entire organism, while the former is a mere snapshot of a certain aspect of it. The only way "pneumonia" is represented in it is *in the act of perception*, that is to say, when it is perceived by an agent whose cognitive capacities allow her to read the idea of pneumonia into the attributes of the image. Furthermore, as an independent entity in its own right, the data-carrier supplies the algorithm information not about pneumonia, but about *itself*, i.e. about its every pixel.

The second node in the machine signification triangle is the *label* (L). It is a word or expression used, for example, to mark the folder which contains a set of data-carriers, which based on the judgement of model designers (or an expert in the field) depict certain instances. This is a category that grasps the result of categorization of the phenomena by a human. As a result, it is tempting to see labels as equivalent to concepts. If that were true, labels would anchor images in reality: one could argue that despite data-carriers being ontologically distinct from the entities they are designed to represent, the fact that images are labelled contributes to the algorithm's ability to successfully differentiate between the localization of inflammation, characteristic to pneumonia, and everything else in the images. Concepts are essentially vessels for ideas; they perform certain functions in the communication, learning and reasoning processes of the agents carrying out these processes. Unfortunately, labels do not work as concepts in the epistemic process of algorithms. They do not convey ideas which would then do their epistemic work in the process: such as, being understood and taken into account in order to identify a new entity. Labels, for the algorithms, are epistemic borders: they demarcate one set of data-carriers from another. Thus, in order to avoid further confusion, one should not equate the label "pneumonia" with the concept pneumonia. We rather should talk about L-***X***ness, as in L-***d***ness (to refer to a pattern derived from the set of images labeled "***d***ogs") and L-***pm***ness (to refer to a pattern derived from the images labelled "***p***neu**m**onia"). Furthermore, from the point of view of the production of machine concepts, labels are a way to introduce to the algorithm the "correct answer", the target towards which it should work and the internal reference point, against which it must calibrate it's calculation.

The third and most important node of the machine signification triangle is *the pattern*. Pattern is a weighted distribution of features across a set of data-carriers, demarcated by a label. This is how it is being identified by the algorithm. The general idea here is that the algorithm scans each data-carrier clustered under each label (such as "pneumonia") on a spatial basis, pixel by pixel (detailed level), segment by segment (aggregated level). Pixels are elementary spatial units of a digital image (akin to a dot). What a human understands as a shape is, in machine epistemology, nothing more than a combination of values of shades of different colors. So the goal of scanning is to calculate, on average, to which degree of intensity of color of each pixel should be present so that a given data-carrier could score as close to 1 as possible, where 1 is preassigned the value of total fit, while 0 is a value of a complete misfit. The result of this calculation is something like a map of averaged weight distribution of features (we use features here in a general sense of a property of a pixel or a segment). It is important that this is *a target oriented process*: it is designed to work towards the correct outcome. The correct outcome is confirming the membership of each data-carrier from the data-set under the label L in the dataset with the label L. Crucial in the process is adjusting the degrees on each segment (aka "weights") so that they are *representative of all the samples in the data set*. Another crucial element in this process is the idea of feedback-loops. A part of algorithm is working backwards: it compares its result with the target value of 1 (this, for us, means something like this: this data-carrier is correctly correlated to the "label L") and, then, given the discrepancies, adjusts (recalculates) the weights distribution map. So essentially the algorithm is tasked with finding a combination of weights that is needed to justify an attribution of a label. The predetermined membership of a data-carrier in the set L is the internal standard of success of the pattern calculation. There is no discovering of patterns that will separate things or events ontologically; the algorithm is creating patterns to match labels.

It is tempting to identify patterns with meaning: one could argue that the result of the features weight distribution amounts to discovering what it means for a data-carrier to be L. But that does not stand. A meaning of a concept, such as of "pneumonia", is such that it must apply to any possible instance of phenomena that one may encounter, while the pattern of features that guarantees the membership of N amount of X-rays in the set labeled "Pneumonia" is such that it applies *reliably* only to *that specific set*. In the end, "85% x" is able to signify anything at all because it is generating a pattern in the distribution of data in a designated set of data-carriers.

Patterns: established (rather than discovered) and projected (rather than predicted)

While human learning, in relevant situations, is about *meaning* (its production and application), DNN's data processing is about *pattern, its creation and projection to a new case*. We suggest that stages of the DNN algorithms that are involved in the production of "85% pneumonia" are better described as: *pattern calculation/calibration* (reflecting the *feature weighting processes)* and *the projection stage*. This terminology grasps more precisely the nature of the processes that happen at each of the stages.

So instead of [γ] we have this schema:

> scan the totality of features in a data set → weigh the distribution of features in this data set → project to a new data-carrier

What is at stake here is the difference between *discovering a pattern* and *creating a pattern*. When it comes to constructing a concept of C or generalizing a principle according to which C occurs, we are dealing with discovering a pattern, which presupposes, among other things, that

- case 1, case 2, case 3, …, case N, that you are observing, share a priori a combination of features and you only need to find it, and that
- by referring to this combination, you will be able to say with certainty *when something is not C*.

Firstly, "discovering" a pattern, among other things, presupposes the possibility of failing to discover it, if the relevant combination of features is not present. Secondly, the features that you single out as belonging to the pattern must be representative of C and *only* of C. However, neither is the case in DNN pattern recognition. Essentially, the DNN is not discovering but creating the pattern. With respect to the first parameter, the DNN cannot fail to establish a pattern; it will always calculate *some* pattern. But this may not be the pattern that correctly relates to the specific entity corresponding to C. With respect to the second parameter, the features that the algorithm singles out are not representative of C. Instead, they are representative only of the set of analyzed data-carries, whose relationship with C is itself questionable. Moreover, the epistemic noise (the features irrelevant to C altogether, such as the background) makes it impossible for these features to be exclusively representative of C.

The pattern calculation is carried out in preparation for *the projection stage* of the AI model. Here DNN is introduced to a data-carrier which is not a part of the original data-carrier sets. The new data-carrier is unlabeled, but the expected output of the model is forced to contain a reference to the labels of the sets of data, for which it has calculated/established the pattern. The algorithm analyzes new data-carriers through the prism of this pattern. More specifically, it evaluates numerically ("weighs") the degree to which each of its features is represented by the pattern, i.e. it calculates for each shared feature of the labelled data-carriers the extent to which its commonality within that data set is represented by the new data-carrier. The system essentially projects (or reads) into a new data carrier the combination of elements typical to data-carriers that are alien to the unknown instance in question. This in effect means that a pattern is forced on new data-carriers. This is especially obvious in the cases where algorithms produce all sorts of bizarre images, where buildings are read into plants, or dogs are projected into any random image. This aspect is exploited in AI-generated art (for an example see https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html), and rightfully so, since the closed analogy with the human cognitive skill would be the free association characteristic to imagination. What this tells us is that the task that the AI system is carrying out is fundamentally different from finding out whether an entity before you is, in fact, a dog, building, or the case of pneumonia. The signification at this stage happens within the following triangle:

> distribution of features (DF) in the set S ↔ (new data-carrier X) ↔ the degree of DF(s) in X

Proper distinction between these two stages in DNN is crucial, on the one hand, because they have different functions in the machine epistemology and, as a result, are open to different types of mistakes. It is more so, because *the machine inference from the former to the latter* (which is in itself the process by which a new phenomenon is categorized) is *the most epistemologically vulnerable one,* i.e. is most prone to errors. The final output of the system cannot be taken in isolation from either of them, since the very meaning of "85% x" is dependent on the specificity of the elements in the data-set the model has been calibrated on [the degree of DF(s) in X]. Despite this, the standard AI narrative downplays the importance of the first stage. This elevates the risk that interrelation of the two segments will be overlooked and overlooked, most importantly, by those who matter the most for the type of decision making we are concerned with here—professional users. This may happen—and often does—because professional users do not come into immediate contact with the first segment of the model, where the calibration happens. This is something developers of the model are dealing with. It is often assumed that this segment of the AI system has no effect on the judgment of the user, only the second segment does. This assumption increases the epistemological vulnerability of the hybrid decision process because it precludes fully informed expert evaluation of the model's output.

Such an evaluation should take into account the following considerations. Firstly, since the feature weighting stage of DNN (where the signification forms) is designed to correlate labels with data (which, for a human observer, translates into the correlation between labels and data-carriers such as images), we are only justified to conclude that an algorithm establishes *the degree of relevance of a certain feature of an image (and not of an entity*) to a certain label (and not the concept of pneumonia). As a result, machine concepts such as "85% pneumonia" do not express or establish a relationship between an entity (such as pneumonia) and an image (such as X-rays) that a medical professional is consulting an AI algorithm about. This is significant, because (a) a data-carrier such as a picture is always a partial representation which opens up a possibility for a wide range of biased judgments; (b) labelling introduces an element of human epistemology into the machine epistemology, which will play a crucial role in the way the algorithm will classify new images. Special attention must be paid to the way the cases are selected and labelled for the pattern calibrating stage of the model. Referring to labels of the original dataset in the explanation of a decision is a hypothetical judgement, not in itself a proof. Machine concepts create frequency correlation between *any possible feature* of this X-ray image, on the one hand, and *any possible feature* of all the images that the algorithm has been scanning in the feature weighting stage. This means that any AI output is *necessarily* contextual. *It is only meaningful in the context of the data-carriers it has been calibrating the pattern on*. And secondly, despite the appeal of the naïve interpretation, β, an AI algorithm *does not really identify anything at all*.[12] The fact that X happens to be a dog, or a confirmed case of pneumonia, is a contingent thing. With this, we must accept that interpreting "85% pneumonia, 10% cancer, and 5% tuberculosis" as "X is very much like pneumonia" (a concept fitting claim) is not justified.

Does "85% x, 10% y, 5% z" amount to: X is more likely to have occurred?

So, returning back to [α], there is one more question that we need to answer: does "85% pneumonia, 10% cancer, and 5% tuberculosis" equal a claim about the probability of the event that this patient is suffering from pneumonia? We argue that not.

To justify the claim that it does, an AI system would have to work out whether, given the evidence, an event x (pneumonia) occurs 85%, y (cancer) 10% and z (tuberculosis) 5% of the time. However, clarifying this sort of "event

[12] This relates to Sullivan (2019) who problematizes "link uncertainty", i.e. the degree to which the model fails to be connected to the phenomenon in question. Sullivan highlights the need of providing additional empirical evidence that ties the model to the phenomenon in order to make a scientifically sound judgment about the phenomenon in the light of the model's output.

uncertainty" is nowhere to be found in the setup and training of the model. The first thing to remind ourselves is that the concept of "event" is not a part of machine epistemology. But even if we would accept, for the argument's sake, that something like events are singled out by the system, we still run into a number of problems with probabilistic interpretation. The sample of images that have been inputted to the AI algorithm (at the pattern calculation stage) cannot be considered as a frequency distribution of the occurrence of an "event" across a population as no representative cross-section data are inputted for that population, nor as a distribution of "events" across time as there are no time series or temporal patterns inputted. In either case, no such data is inputted that would form the basis of understanding the frequency of occurrence of event x (cases of pneumonia) alone, or of event x conditional on the co-occurrence of event(s) y, z (cases of cancer and tuberculosis). Secondly, even if the above assumption is relaxed,[13] the inputted "events" are not clearly delineated and mutually exclusive, whereas the explanatory and control factors do not clearly and uniquely correspond (or being corresponded by the machine) to, again, regularly occurring, clearly delineated and mutually exclusive "events". Therefore, when interpreting "85% pneumonia, 10% cancer, and 5% tuberculosis" as an equivalent of a claim about the probability of an event, we are effectively asking from the AI model to derive temporal (frequency) patterns where no such patterns have been given to it, therefore misapplying to its % outputs extremely popular frequentist statistics notions into a type of uncertainty that is not of a "how-often-an-event-occurs" nature.

But what *is* worked out by an AI algorithm? The algorithm in fact has been provided with the information and algorithmic machinery to recognize *spatial* rather than *temporal* patterns. What the AI system sees as its input is (a) generally unstructured and not always relevant information contained in a data carrier, clustered according to (b) labels. The algorithmic machinery is then tasked with structuring the provided information; but the expectation is that it would in some way "learn" to reconstruct "events" x, y and z. What is therefore happening is that spatial information is used to train a model to predict how much an arrangement of features is characteristic of x-ness, y-ness and z-ness, without any constraints on order of occurrence or mutual exclusivity.

The naive temporal interpretation, following the frequentist viewpoint of "85 out of 100 cases in the past, this was the signature of x (pneumonia), whereas in 10 cases the signature of y (cancer) and in 5 cases the signature of z (tuberculosis)" does not make sense, since the actual task that the machine performed was to establish in what arrangement certain spatial features make x (presented to it by data-carriers labelled "pneumonia"), y (data-carriers labelled "cancer") and z (data-carriers labelled "tuberculosis") concurrently and in the presence of each other. A closer—but, as we have shown earlier, still not accurate—reading of the output would be that in 85% of these cases the spatial features were arranged in such a manner that were tagged as x, but also as y in 10% and as z in 5% of the cases.

If we now refer the reader back to the beginning of the paper where we asked how warranted is the intuitively appealing mode of hybrid reasoning [α]

> if [α1] an output of an AI algorithm is ("85% pneumonia, 10% cancer, and 5% tuberculosis"), then [α2] it is *most likely* pneumonia, therefore [α3] [for the diagnostic purpose] it is pneumonia.

we can conclude that the move from [α1] to [α2] is problematic, and so [α] is not justified. The probabilistic inference "if an output of an AI algorithm is 85% pneumonia, 15% cancer, 5% tuberculosis then it is most likely a case of pneumonia" is in fact *an assumption*, which requires proof from an independent source. And, thus, the inference about a matter of fact [α3] is a leap of faith, while the practical step of treating the patient from pneumonia cannot be supported. Relying only on the output of the model to corroborate the decision about the patient's treatment may happen to be the right one, but only accidentally so.[14] A responsible role in this hybrid decision making process requires support from other sources.[15]

How to interpret the machine output 85% x, 15% y, 5% z?

So far we have been arguing how not to interpret the machine output. But how are we to interpret it? In this section we make two claims: (1) that "85% x, 15% y, 5% z" only warrants us to make a judgment about trivial similarity between x, y, and z, and (2) that it essentially maps *a range of overlapping boundaries between fuzzy concepts represented by the labels x, y and z,* given the constraints of the data-carriers used to calibrate the model. Let's explain this.

[13] One may claim, for instance, that the model developers did manage to assemble a sample of data-carriers that is representative of the (presumably also clearly defined) population.

[14] Grote and Berens (2020) raise a similar concern, when saying that "whereas involving machine learning might improve the accuracy of medical diagnosis, it comes at the expense of opacity when trying to assess the reliability of given diagnosis." One part of this is opacity (see more, e.g. Carabantes, 2020; Heinrichs & Eickhoff, 2019), but what is of more importance here is the question of reliability.

[15] What sort of evidence and tools are needed for the medical diagnosis should be discussed in the context of the nature of medical diagnosis itself. We won't be able to discuss these issues here, but we would refer the reader to an interesting recent discussion on the matter in relation to AI decision support systems by, e.g., Kudina & de Boer (2021) and van Baalen et al. (2021).

Trivial similarity

The closest equivalent to the machine output in human epistemology is the *judgment about similarity*. Similarity between case 1, case 2, case 3, …, case N is non-trivial when it provides sufficient information for establishing the relatedness of the new case to those previously observed. We are comparing known cases in order to be able to categorize a new case: what is the most informative result of comparison? Since we want to find out whether or not the new case is the same phenomenon as the previously observed instances, arguably, it is a judgement about *the similarity of relevant features*, i.e. about the sharedness of unique distinguishing features between such cases/entities. This is what a trained professional would do.

One way to fail to make such a judgment about the relevant features, would be to conduct a comparison by taking into account

- only *randomly* limited information about the entity in question;
- the totality of features in case 1, case 2, case 3, …, case N, that is, by comparing all features in case 1 to all features in case 2, case 3, …, case N;
- the totality of features of the environment, in which case 1, case 2, case 3, …, case N are situated.

This would run a high risk of picking up irrelevant and accidental elements, which are non-informative. The result would be similarity which does not inform whether the cases are instances of the same thing. We can draw an analogy with a mistake a radiologist may make if he/she takes into account an irrelevant feature of the image, such as the contrast medium injected into the joint.

This is, however, exactly what a DNN does, because

- it extracts information from a partial representation of an entity (from a data-carrier);
- it scans and compares all segments of all data-carriers in a set;
- not all segments of the data-carrier represent the entity in question.

The best way to describe what an AI system does is as: finding *anything that can possibly unite* case 1, case 2, case 3, …, case N. This amounts essentially to *constructing similarity*, rather than identifying or discovering existing similarity of relevant features. The problem with this approach is that in order to be successful in constructing similarity, one does not need to take into account how things are. It all depends on the level of generality, i.e. on where one sets the bar for the abstraction of features. When comparing a dog with an oyster, one may not find similarity on the level of species, but will find these two similar as living entities, vs. non-living ones such as rocks. And in the end, being is the underlying unity of everything, including stones, dogs, trees, and Hegel's concept of the Absolute Spirit. This, however, comes at a price: the more abstract you go, the less informative and more irrelevant the similarity judgments become for the specific decision making purposes.

So, what we are warranted to infer from the AI's output is a statement about *trivial similarity*. The algorithm does not discriminate between any bit of data, i.e. between any detail of the data-carrier when it performs its comparison: empty space, noise, background features, any imperfections in the data-carrier—everything is taken into account. One could object by saying that on the feature weighting stage, the algorithm does exactly that: it extracts the most relevant distinguishing features. But that is simply not true, otherwise we would not have cases when an AI algorithm correctly identifies a wolf in a picture because of snow in the background. Even when the algorithm outputs a higher percentage of wolf-ness in the picture which displays a wolf, this is nothing more than a coincidence. Calibrating the model is normalizing this accidental correctness. The ambiguity of the concept of similarity explains the confusion about the concept of pattern. In a narrow sense of the word, pattern is a combination of features that are representative of a phenomenon; a unique combination of elements that make a certain entity distinguishable from other entities. This explains why we are able to identify entities by a certain pattern. This is not what pattern is in DNN epistemology. Here pattern is used in a wide sense of a shared property, something that is common to any two or more random entities. Everything shares a pattern with everything else. Thus, the bizarre images of AI art when it sees dogs in images of trees is not an error in machine epistemology. They are, however, in human epistemology.

On the fuzziness of machine concepts

"85% x, 10% y, 5% z" is in fact describing degree of trivial similarity with sets x, y, z instead of either x, or y or z when seen as treating concepts; and the amount of times x, y, z was tagged as such given a certain arrangement of features, instead of frequency of (not) being x, y or z in past cases when seen as treating events. Much of the confusion about "85% x, 10% y, 5% z" comes from a lack of understanding about the nature of the transformation of concepts *as epistemic units* within hybrid reasoning, which results from the cooperation of the human with AI that is based on DNN. As was just demonstrated, the DNN establishes trivial similarity between concepts (represented by labels) while calculating the degree to which the distinction between them can be blurred. This is to say that DNN fuzzyfies concepts with no regard to the nature of the concepts that stand behind the original labelling.

In the domain of soft computing, the distinction between hard and fuzzy concepts was best explained by Zadeh (1965, 1975) and is insightful for the explanation of uncertainty. If we describe concepts as sets, then *hard concepts* are those that have sharp boundaries which delineate members of the set from non-members. Membership in such a set is a binary parameter. "Dog" can be approached as such a concept: X is either a dog or not. In cases when it is not clear whether X is a dog or a wolf, the uncertainty would be due to incomplete information. Fuzzy concepts, however, do not impose sharp boundaries and the membership in these sets is a matter of degree. The uncertainty in such cases comes from blurred boundaries and the fact that something can belong to multiple sets concurrently, but at varying degrees. A widely used example of a fuzzy concept is temperature, where hot and cold do not share a crisp boundary but rather blend into each other along a continuum of varying degrees of hotness and coldness. Due to the way they generate signification, machine concepts (such as 85% x) do not and cannot generate a function for binary membership. They always determine a degree to which X can be seen as belonging to a certain labelled set. Machine concepts of the type generated by DNN are necessarily fuzzy. In the case of the label "dog", the concept enters the hybrid reasoning system as a hard concept because it is a hard concept in human epistemology. After the machine output, the same label represents a fuzzy concept of the dog, because the algorithm has tried to establish a degree of similarity of a certain arrangement of spatial features to those shared by the members of the data set labelled "dog", i.e. to L-*d*ness. The machine output "85% x, 10% y, 5% z" establishes similarity between a predetermined selection of machine-concepts; it is *mapping a range of overlapping boundaries between concepts represented by the labels x, y and z*. For instance, if x, y and z would refer to labels "dog", "wolf" and "cat", their respective percentages in model output would communicate the concurrent degree of membership of a given unlabeled data-carrier, DC, in the fuzzy sets labelled "dog", "wolf" and "cat". This, however, does not translate to "more like"/"less like a dog" but to "there is x amount of L-*d*ness, y amount of L-*w*ness, and z amount of L-*c*ness in your specimen DC.

Rather than following the event-probabilistic interpretations of judgements about fuzzy concepts, it seems more appropriate to adopt Zadeh's concept of *possibility*.[16] In a decision making context, membership degree x(/y/z) of DC in a fuzzy set X(/Y/Z) has been re-defined by Zadeh (e.g. 1999) as degrees of possibility that DC is X/Y/Z and the model output "85% x, 10% y, 5% z" should be rather seen as answering the question about how *possible* it is that what we are faced with is X and not a question about how *probable* it is.[17] The epistemological contribution of model output in the decision making chain sketched in Fig. 1 is therefore better understood as reflecting the possibility that the entity is X/Y/Z. We won't be able to go into possibility theory in any detail here, but it is important to draw attention to a couple of considerations. First, as Zadeh notes: "a high degree of possibility does not imply a high degree of probability, nor does a low degree of probability imply a low degree of possibility" (1999, p. 14). To use Zadeh's own example, while it is quite possible that Hans will eat 3 eggs for breakfast (possibility = 1), it is not very probable that he will do so (probability = 0.1). This is because the criteria that determine whether a phenomenon is more possible are different from the criteria that determine whether it is more probable. Probability is a question of frequency of a phenomenon having a certain characteristic across a set of the phenomenon's occurrences. Possibility distribution, on the other hand, is a function of fuzzy restrictions associated with a concept C. What is at stake here are "the degrees of compatibility", which can otherwise be explained as the "degrees of ease" with which a certain characteristic (represented by C) can be attributed to the phenomenon in question. So, while "85% pneumonia, 10% cancer and 5% tuberculosis" may point to the degrees of ease with which, given the context, one would ascribe to the observed case the respective concepts, that cannot be translated to the probability distribution of the phenomena expressed by these concepts. This brings us to the second point. What contributes to the "degree of ease" (in the case of possibility) is a variety of highly contextual (and often subjective) determinants. In the case of Hans eating eggs, this determinant is the ease with which he can eat 3 eggs (and, say, not experience indigestion). While in the case of a judgment like "John is young", the possibility of John being young is a matter of ease with which in a certain cultural framework people attribute "young" to various age

[16] An insightful discussion on the fundamentally different kinds of uncertainty dealt with by sciences and engineering, entitled 'The elusive concept of uncertainty', took place in UC Berkeley's BISC group mailing list between Lotfi Zadeh and other prominent researchers in AI and soft computing. There, among others, Zadeh (27.11.2013) outlined three main types of uncertainty: relating to repetitive random events; relating to incompleteness of information on singular or unique events; relating to un-sharpness or fuzziness of class boundaries (possibility theory). In response, Marianne Belis (11.5.2014) re-grouped uncertainties, depending on the sharpness of the boundaries, into two types: due to the incompleteness of information (for concepts with sharp or "hard" boundaries) and due to fuzziness (un-sharpness) of the class boundaries. Probability is a concept suitable for the first type; while possibility grasps best the second type.

[17] In this respect, note Sullivan (2019) who claims that models such as DNN contribute to "how-possibly explanations", i.e. explanations about possible correlations and dependencies, and not "how-actually" or "why-questions" about the actual phenomenon.

categories, and John’s age as attributes to these age categories. A lot of this has to do with the use of language and conventions. In the case of AI systems, it is the DNN itself that creates such a determinant in application to the fuzzy sets of machine concepts that it creates. Whether such a determinant coincides with the constraints on attribution of the respective concepts from the human epistemology, is an open question.

Conclusion: normative constraints on inferences from AI output in decision making

We started with the question: in the situation when we do not know for certain what an entity is, and given the output of an AI decision support system,“85% x, 10% y, 5% z”, are we warranted to jump to the conclusion that the entity in question is *most likely* x? To answer this, we had to take a closer look at the processes that underlie machine epistemology and to understand what the unwarranted interpretations of the machine output are. This required analyzing machine concepts and distinguishing them from concepts in human epistemology, as well as evaluating the applicability of the event prediction interpretation. This led to a revised understanding of the ways the elements of the machine epistemology should be interpreted into the hybrid (human/machine) decision making processes.

More specifically, we wanted to understand if, when the algorithm processes a data-carrier (in our example an X-ray), DC, we can move from the machine statement DC = {x: %, y: %, z: %} to a conclusion that, for the diagnostic and treatment purposes, the patient’s condition is indeed x. Is AI output by itself enough to warrant this move? And if not, then what else is needed? We argued that the conclusion that the patient is indeed suffering from x is unwarranted, both when model output is seen as concerning the concept x and event x. We have shown that, because the AI model is equipped to answer a different kind of uncertainty than what the human interpreter has in mind when reading the output. We argued that these significations are nowhere to be found in the algorithmic machinery and information available to an AI model. We concluded that model output rather represents degrees of membership of the data-carrier DC in a number of fuzzy sets, delineated by labels (L-*x*ness). Moreover, in a decision making context we suggested that DC = {x: 85%, y: 10%, z: 5%} is better understood as representing the possibility of DC being x/y/z and that additional information is needed to answer something about the respective probability. In other words, one cannot rely solely on the output of the model to corroborate a decision about the patient’s diagnosis and treatment.

Of course, if we did have a strong AI,[18] such that it would be able to navigate in contexts not accessible to the human mind, that is to say, if we had an AI that were able to process the mass of its own perceptions of reality in its full totality, then it could be closer to generating epistemic processes like those of human induction/deduction. And if it were able to extract relevant elements from the mass of perceptions, then it would be useful in mapping a range of fuzzy sets, representative of all relevant alternatives, and warranting further questions about probability. But the reality is that we do not have such an AI.

Ethical implementation of AI decision support systems in such high-risk areas as health care requires a responsible stance on behalf of the human decision maker who is involved in the hybrid reasoning process. This, in turn, creates a number of responsibilities for system developers and marketers as well. More specifically:

(a) The human decision maker has *an epistemological responsibility* to develop a factually informed understanding of the processes that lead to the formation of AI’s recommendation. The developers and marketers have a corresponding responsibility to provide realistic description of the processes involved, and not to deceive potential (professional) users concerning the capacities of the system and its role in the decision making process. This includes avoiding fundamental mistakes in the utilization of pattern recognition technology, such as the belief that the machine has the ability to recognize entities from a sensory input.

(b) The human decision maker has *an epistemic responsibility* of being aware about the limitations of the particular AI decision support system he/is using. This is required in order to prevent situations when the output of the model is not-applicable to the specific circumstances of the patient. This creates a corresponding requirement for the developers and marketers to supply the model with a manual which includes an accurate description of the model, detailed description of the data set it was calibrated on, and the limitations that follow from this calibration as to the new cases that the model should be processing. The AI’s applicability range must always be clearly defined, and the users must be warned that it can never be used a-contextually, however narrow that context might be.

(c) The human decision maker should avoid over-relying on the AI model. Its output should always be placed within the context of other relevant diagnostic procedures. As diagnosis is itself a fuzzy set, the output

[18] In this context, this term refers to an artificial system capable of thought, consciousness and emotions in a genuinely human way (Searle 1980).

must always be supplemented by other tools that give a better understanding of the relevant causal processes. The developers and marketers should warn the potential professional users about this necessity.

Funding: Not applicable.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors report no conflicts of interest.

References

Almalki, Y. E., Qayyum, A., Irfan, M., Haider, N., Glowacz, A., Alshehri, F. M., Alduraibi, S. K., Alshamrani, K., Alkhalik Basha, M. A., Alduraibi, A., Saeed, M. K., & Rahman, S. (2021). A novel method for COVID-19 diagnosis using artificial intelligence in chest X-ray images. *Healthcare, 9*(5), 522. https://doi.org/10.3390/healthcare9050522

Almaslukh, B. (2021). A Lightweight deep learning-based pneumonia detection approach for energy-efficient medical systems. *Wireless Communications and Mobile Computing*. https://doi.org/10.1155/2021/5556635

Badré, A., Zhang, L., Muchero, W., et al. (2021). Deep neural network improves the estimation of polygenic risk scores for breast cancer. *Journal Human Genetics, 66*, 359–369. https://doi.org/10.1038/s10038-020-00832-7

Belis, M. (2007). The causal roots of probability. In F. Russo & J. Williamson (Eds.), *Causality and probability in the sciences.* College Publications.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *Npj Digital Medicine, 3*, 118. https://doi.org/10.1038/s41746-020-00324-0

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*. https://doi.org/10.1016/B978-0-12-818438-7.00002-2

Boon, M. (2020). How scientists are brought back into science—the error of empiricism. In M. Bertolaso & F. Sterpetti (Eds.), *A critical reflection on automated science. Human perspectives in health sciences and technology.* (Vol. 1). Springer. https://doi.org/10.1007/978-3-030-25001-0_4

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI &amp; Society, 35*, 309–317. https://doi.org/10.1007/s00146-019-00888-w

Chollet, F. (2018). *Deep learning with python*. Manning.

Elder, A., Ring, C., Heitmiller, K., Gabriel, Z., & Saedi, N. (2021). The role of artificial intelligence in cosmetic dermatology—current, upcoming, and future trends. *Journal of Cosmetic Dermatology, 20*, 48–52. https://doi.org/10.1111/jocd.13797

Fujita, H. (2020). AI-based computer-aided diagnosis (AI-CAD): The latest review to read first. *Radiological Physics and Technology, 13*(1), 6–19.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205–211.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402–2410.

Heinrichs, B., & Eickhoff, S. B. (2019). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping, 41*(6), 1435–1444. https://doi.org/10.1002/hbm.24886

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., & Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 590–597).

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*. https://doi.org/10.1287/isre.2020.0980

Kaplan, A., Cao, H., FitzGerald, J. M., Iannotti, N., Yang, E., Kocks, J. W. H., Kostikas, K., Price, D., Reddel, H. K., Tsiligianni, I., Vogelmeier, C. F., Pfister, P., & Mastoridis, P. (2021). Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. *The Journal of Allergy and Clinical Immunology: In Practice, 9*(6), 2255–2261. https://doi.org/10.1016/j.jaip.2021.02.014

Kapoor, R., Walters, S. P., & Al-Aswad, L. A. (2019). The current state of artificial intelligence in ophthalmology. *Survey of Ophthalmology, 64*(2), 233–240.

Kudina, O., & de Boer, B. (2021). Co-designing diagnosis: Towards a responsible integration of machine learning desicion-support systems in medical diagnostics. *Journal of Evaluation in Clinical Practice, 27*(3), 529–536.

Liu, Y., Zhong, X., Cheng, J., Pi, Y., Cai, H., Jiang, L., Yang, P., Xiang, Y., Jianan, W., Li, L., Yi, Z., & Zhao, Z. (2021). Automatic rapid identification of malignant carcinoma bone metastatic lesions by deep neural network based artificial intelligence. *Journal of Nuclear Medicine, 62*(1), 1180.

Mayo, R. C., Kent, D., Sen, L. C., Kapoor, M., Leung, J. W. T., & Watanabe, A. T. (2019). Reduction of false-positive markings on mammograms: A retrospective comparison study using artificial intelligence-base CAD. *Journal of Digital Imaging, 32*(4), 618–624.

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science &amp; Medicine, 260*, 113172.

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning: A study of the influence of language upon thought and on the science of symbolism*. Harvest.

Peirce, C. S. (1998). *The essential peirce. Peirce edition project* (Vol. 2). Indiana University Press.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering, 2*(3), 158–164.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint. arXiv:1711.05225.

Rundle, C. W., Hollingsworth, P., & Dellavalle, R. P. (2021). Artificial intelligence in dermatology. *Clinics in Dermatology*. https://doi.org/10.1016/j.clindermatol.2021.03.011

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417–457.

Sheth, D., & Giger, M. L. (2020). Artificial intelligence in the interpretation of breast cancer on MRI. *Journal of Magnetic Resonance Imaging, 51*(5), 1310–1324.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035

Ting, D. S., Foo, V. H., Yang, L. W., Sia, J. T., Ang, M., Lin, H., Chodosh, J., Mehta, J. S., & Ting, D. S. (2021). Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology. *British Journal of Ophthalmology., 105*(2), 158–168.

Topol, E. J. (2019). *Deep medicine. How artificial intelligence can make healthcare human again*. Basic Books.

van Baalen, S., Boon, M., & Verhoef, P. (2021). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice, 27*(3), 520–528.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Pysarenko, H. T. K., et al. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transanctions on Medival Imaging*. https://doi.org/10.1109/TMI.2019.2945514

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences, 8*(3), 199–249.

Zadeh, L. A. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems, 100*, 9–34.