Ethics and Information Technology (2022) 24:26
https://doi.org/10.1007/s10676-022-09650-1

ORIGINAL PAPER

Karl Jaspers and artificial neural nets: on the relation of explaining and understanding artificial intelligence in medicine

Georg Starke[1,2] · Christopher Poppe[1]

Abstract
Assistive systems based on Artificial Intelligence (AI) are bound to reshape decision-making in all areas of society. One of the most intricate challenges arising from their implementation in high-stakes environments such as medicine concerns their frequently unsatisfying levels of explainability, especially in the guise of the so-called black-box problem: highly successful models based on deep learning seem to be inherently opaque, resisting comprehensive explanations. This may explain why some scholars claim that research should focus on rendering AI systems understandable, rather than explainable. Yet, there is a grave lack of agreement concerning these terms in much of the literature on AI. We argue that the seminal distinction made by the philosopher and physician Karl Jaspers between different types of explaining and understanding in psychopathology can be used to promote greater conceptual clarity in the context of Machine Learning (ML). Following Jaspers, we claim that explaining and understanding constitute multi-faceted epistemic approaches that should not be seen as mutually exclusive, but rather as complementary ones as in and of themselves they are necessarily limited. Drawing on the famous example of Watson for Oncology we highlight how Jaspers' methodology translates to the case of medical AI. Classical considerations from the philosophy of psychiatry can therefore inform a debate at the centre of current AI ethics, which in turn may be crucial for a successful implementation of ethically and legally sound AI in medicine.

The promises of artificial intelligence for medicine

The integration of artificial intelligence (AI) seems bound to reshape the practice of medicine (Topol, 2019). Due to the convergence of Big Data, increased computational capacities and the rise of deep learning, a new generation of AI systems promises vast improvements, from new research approaches to their clinical implementation at the bedside. While for some authors the current hype of AI creates a danger of bringing about a new *AI winter*, i.e., a period of decreased interest and funding (Müller, 2020; Floridi, 2020), the underlying technology may still usher in an age of *Deep Medicine*, given its tangible successes (Topol, 2019). After all, AI can provide tools that improve clinical outcomes across disparate medical specialties, from dermatology (Esteva et al., 2017) to pathology (Campanella et al., 2019), from intensive care (Hyland et al., 2020) to plastic surgery (Knoops et al., 2019) and psychiatry (Bzdok & Meyer-Lindenberg, 2018). Questions concerning the ethical and responsible design and use of medical AI are thus of high urgency and importance.

One major challenge to the implementation of AI in high-risk settings such as medicine lies in the lack of explainability of many current AI systems in healthcare (Vayena et al., 2018; Amann et al., 2020). This challenge results from the opacity of AI models, which in particular deep learning models exhibit (Burrell, 2016). Explainability seems of crucial instrumental value to foster trust in AI systems, to

✉ Georg Starke
georg.starke@unibas.ch

[1] Institute for Biomedical Ethics, University of Basel, Bernoullistr. 28, 4056 Basel, Switzerland

[2] Intelligent Systems Ethics Group, College of Humanities, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Published online: 16 June 2022

correct a model's errors and to enable vital ethical aspirations like informed consent. Accordingly, ethical guidelines for the implementation of AI have even granted explainability or the related ideal of explicability a place alongside the four influential principles of biomedical ethics by Beauchamp and Childress, complementing beneficence, non-maleficence, respect for autonomy and justice (Floridi et al., 2018; Beauchamp & Childress, 2019). In addition, as the European General Data Protection Regulation (GDPR) highlights, explainability does not constitute a mere ethical recommendation but has become a legal requirement in some jurisdictions and is seen as a part of fundamental rights (Wachter et al., 2017).

Yet, despite its importance, exact, formal definitions of explainability are scarce and often differ across research domains (Adadi & Berrada, 2018). Mittelstadt and colleagues (2019) and Durán (2021) have examined the notion of explainability cautiously with regard to the philosophy of science, situating it in the broader context of scientific explanations. However, as Páez (2019) has convincingly argued, explanations resting on full model transparency which would allow to answer counterfactual questions run into severe and potentially insurmountable problems. Hence, the complexity of a model renders certain types of AI inherently opaque to causal explanations. While this may not preclude epistemically more modest explanations for specific, single decisions of an ML system, it still seems worth turning to a scientific tradition that has long struggled with the problem of explaining phenomena that defy full mechanistic explanation, namely philosophy of psychiatry.[1] In particular, we argue that Karl Jaspers' seminal framework of explaining and understanding in psychopathology provides a rich conceptual background that can be fruitfully adapted to address the challenges posed by current AI systems developed for medical purposes.

Our argument proceeds in five steps. First, we provide a short primer on current debates about the explainability of AI, highlighting its limits. Second, we turn to Jaspers, elaborating the elements of his theoretical framework for the debate at hand. In a third step, we argue why psychopathology can serve as a model to develop a framework of explaining and understanding AI, and fourth, why applying a model from psychopathology to AI is warranted, despite the danger of anthropomorphism. Finally, bringing together these considerations, we suggest a framework for understanding and explaining medical AI inspired by Jaspers. We conclude by drawing on examples of medical AI to highlight the practical and ethical implications of our approach.

[1] In the same vein, Páez (2019) also turns to a distinction derived from psychology between functional and mechanistic understanding to advance his argument (Lombrozo & Gwynne, 2014).

The challenge of explainable AI systems in medicine

Rendering AI systems explainable is commonly regarded as crucial for their successful implementation. Consequently, the development of explainable AI (XAI) takes centre stage in myriads of research efforts worldwide (Adadi & Berrada, 2018). Explainability has the instrumental value enabling crucial epistemic and ethical goals (Floridi et al., 2018). On the epistemic side, by allowing closer scrutiny of a system's decisions, XAI promises developers, regulators and end users the possibility to spot systematic mistakes, correct erroneous decisions and improve the system's performance. In turn, these properties promote important ethical aims, such as fostering informed consent, accountability and avoiding discriminatory biases.

In clinical settings, the degree of a system's explainability may also have important consequences for the complex web of relations between software developers, regulatory bodies, physicians, and patients (Amann et al., 2020). For example, explainability is not only crucial for obtaining informed consent, which requires at least some minimal standards of knowledge, but is also a vital property for promoting trust in a specific system (Diprose et al., 2020). Furthermore, from the perspective of patients, some degree of explainability is required to be able to contest an AI's diagnostic decision – an important ethical desideratum, rooted in the patients' right to defend themselves against harm (Ploug & Holm, 2020).

Unfortunately, the opacity of AI systems often resists simple explanations. Besides intentionally created secrecy measures within a program, opacity can come in the guise of technical illiteracy on the side of its users or as a system's property, necessarily following from its design and use (Burrell, 2016). Here, we are only interested in the latter. Such necessary opacity, commonly addressed as black-box problem in AI ethics, is particularly prevalent in deep learning models based on artificial neural nets (ANN). To some extent, this opacity may constitute a necessary characteristic of the program, following directly from an architecture with multiple hidden layers and a huge number of weights, optimized with vast and complex training data containing multiple features.

At the moment, approaches to increase an AI system's explainability often focus on visualizations, providing e.g. a heat or saliency map for a program's decision. As Mittelstadt and colleagues (2019) have succinctly pointed out though, such approaches fall short of common human expectations towards a meaningful explanation, characterized by their contrastive, social, and selective nature. In the same vein, Páez (2019) has argued in favour of a pragmatic turn that cedes unrealistic attempts aiming at full causal

explainability in favour of interpretative models that are easily accessible to the intended users. Within the specific context of medicine, Alex London has famously taken an even more provocative approach by arguing that we should prioritize the diagnostic or predictive accuracy of an AI system over its explainability (London, 2019). Similarly, we also agree with the view advocated by Durán & Jongsma (2021) that reliable, yet opaque black box algorithms can provide trustworthy tools for improving medical care.

Yet, given the ethical and epistemic importance of explainability, it would seem prudent to aim for a framework that retains the important aspirations ingrained in the project of rendering medical AI explainable wherever possible. As in other ML systems, explainability would comprise both ex-ante considerations, that focus on the input to a particular program, and ex-post evaluations, scrutinizing the output of a trained algorithm (Braun et al., 2021). Furthermore, in the specific context of medicine, explainability will also need to take into account the complex relation between physician, patient, and ML system, e.g., because physicians need to explain a decision to their patients (Braun et al., 2021). To enable successful forms of such communication and thereby establish the necessary preconditions for trust in a particular program, it will, as argued elsewhere, be crucial to not merely disclose information but render them intelligible, accessible, and assessable to the concerned parties (Starke, 2021; Arbelaez Ossa et al., 2022) .

These theoretical considerations are also supported empirically, e.g., by a recent survey among 170 physicians in New Zealand which confirmed that physicians' understanding of a ML model, their ability to explain the program's output to their patients and their trust in using it are indeed related to each other (Diprose et al., 2020). In light of these findings, it seems advisable to address the particular challenges of medical ML through a lens which not only discerns between different notions of explaining and understanding but relates them to each other in a systematic manner. As we will show in the following, Karl Jaspers' methodological groundworks in psychopathology offers this very kind of framework.

Karl Jaspers: explaining and understanding

In his seminal *Allgemeine Psychopathologie* (AP) from 1913 (cited in the 4th edition; Jaspers 1946), Jaspers famously distinguished between different approaches to address the epistemic difficulties of dealing with the inner life of his patients. Crucial to his writings is the distinction between explaining and understanding. This classic distinction drew on debates about methodological differences between humanities and natural sciences spearheaded by the German philosopher Wilhelm Dilthey in the late 19th century, who famously declared: "Nature we explain, but psychic life we understand" (1894, p. 144, quoted in Kumazaki 2013). It also relates to Wilhelm Windelband's distinction between "nomothetic" and "idiographic" empirical sciences, with the former seeking "the general in the form of a law of nature", and the latter seeking "the particular in the form of the historically defined structure" (Windelband, 1980 [1894], p. 175).

Expanding on this framework, Jaspers developed a systematic approach encompassing a multi-faceted attempt to integrate subjective and objective phenomena and inferences, which comprised three consecutive steps. According to Jaspers, any attempt of explaining or understanding first needs to fully grasp the relevant facts (Jaspers, 1913, p. 22 f.), that encompass both objective and subjective data. For an objective psychopathological assessment, the evaluation draws on outward observations and quantifiable data such as persons' interaction with their environment or their quantifiable performance in memory assessment (Jaspers, 1946, p. 130). Ideally, such objective assessment would imply that the clinician refrains from all theoretical and personal prejudices and presuppositions, relying for example on objective measures such as established psychometric scales, allowing for interindividual comparisons. In contrast, to take stock of the subjective facts of the inner life of a person such as their lived experience of a delusion, Jaspers suggests a 'phenomenological' approach, loosely based on Edmund Husserl's phenomenology, attempting to grasp an individual's own perspective of their lived experience. As Jaspers describes the method with regard to patients in his psychopathology:

"The task of phenomenology is to visualize the mental states that the sick really experience, to look at them according to their relationship, to limit them as sharply as possible, to distinguish between them and to assign them fixed terms." (Jaspers, 1946, p. 47) [2]

Jaspers himself calls this phenomenological realization and envisionment of a psychological state "static understanding" (Jaspers, 1946, p. 24). It should be noted though that this is not the kind of understanding in which we are interested here.

Having taken stock of the 'factual data', the psychopathologist then needs to make sense of these fragmentary data by investigating the relations between them (Jaspers, 1946, p. 23). Jaspers proposes two ways, and it is here that we finally encounter the distinction between understanding ("verstehende Psychologie") and explaining ("erklärende Psychologie") that is of interest to our argument.

"We need to draw a distinction between these relations that is just as fundamental as the distinction between

[2] Here as in the following, translation from the German original is provided by the authors.

subjective psychopathology (phenomenology) and objective psychopathology. 1. By putting ourselves into the psychic situation, we *understand genetically* how one psychic event emerges from another. 2. By objectively linking several factual data into regularities based on repeated experiences, *we explain causally*." (Jaspers, 1946, p. 250).

For Jaspers, *explaining* therefore hinges on identifying a clear causal connection between cause and effect, and is commonly rooted in biology. According to Jaspers, establishing such an *explanatory* relation allows to formulate a rule that is valid for similar instances (Jaspers, 1946, p. 251). Such explanations therefore closely correspond to the methodological approach of the natural sciences, which according to Jaspers only investigate genuine causal relations (ibid.).[3]

In contrast, and going beyond the scope of natural science, subjective *understanding* concerns itself with comprehensible, meaningful relations that are related to personality and biography. It establishes meaningful connections by drawing on the psychopathologist's own inner experiences, resulting in a "direct evidence that we cannot trace back any further" (Jaspers, 1946, p. 252). The evidence of these understandable relations is not based on genuine causal explanations but rather on psychological plausibility, and is achieved by contemplating mental life (Jaspers, 1946, p. 48). Jaspers calls such understanding "genetic", to distinguish it from the "static understanding" mentioned above (Jaspers, 1946, p. 252). Since we are only interested in this form of understanding, we will omit the qualification as "genetic" in the following.

In a nutshell, Jaspers proposes a model of psychopathology that offers a subjective and an objective approach both for the gathering of factual data and for establishing meaningful relations between them. The psychopathologist first needs to gather all relevant observations from their patient, including the patient's subjective mental state as well as their objective environment and biological state. Having brought both together in a full description, there are then two ways to establish meaningful relations between them, either through subjective ("genetic") understanding or objective explanations. A schematic depiction of the complementary subjective and objective approaches is provided in Fig. 1, to give a succinct overview over Jaspers' terminology.

Jaspers' model has been subject to fundamental criticism, including a recent call to give up the distinction between understanding and explaining in psychiatry altogether (Gough, 2021). More important to our argument, however, are attempts to disentangle the notion of causality in the context of Jaspers' distinction. For instance, many current scientific claims would possibly not fall under Jaspers' rigorous definition of explainability, as long as causal relations remain unclear.[4] Drawing on the writing of Elizabeth Anscombe, Hoerl (2013) has therefore suggested to describe both explaining and understanding in terms of causality, but with an important difference: Explaining provides "general causal claims linking types of events", whereas understanding "is concerned with singular causation […] – i.e. with the particular way in which one psychic event emerges from or arises out of another on a particular occasion." (Hoerl, 2013, p. 111).

This reading, distinguishing between general causal claims and singular causation, in fact mirrors Jaspers' own distinction between two different kinds of causality that seems in line with Husserl's distinction of volitional and natural causality (Spano, 2021; Husserl, 2020), yet sometimes renders Jaspers' arguments seemingly contradictory. While causal relations in the strict sense are, according to Jaspers, only to be found in the objectifiable outward observations of the natural sciences, (Jaspers, 1946, p. 250) he sometimes also employs a notion of causality that grasps the understandable subjective phenomena:

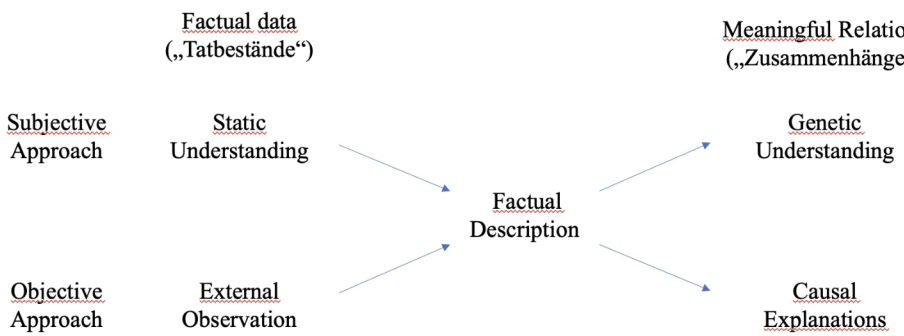

Fig. 1 Schematic representation of the subjective and objective evaluation in Jaspers' psychopathological approach

[3] It should be noted that Jaspers' original model from 1913 predates the vast philosophical debates concerning scientific explanations that take their cue from Carl Hempel's Deductive-Nomological Model from 1942.

[4] We would like to thank one of our anonymous reviewers for pointing this out.

One has also called the intelligible connections of the mental *causality from within*, and thus denoted the unbridgeable abyss that exists between these merely parabolically causal connections and the genuine causal connections, the *causality from without*." (Jaspers, 1946, p. 250).

If we follow this distinction by Jaspers and Hoerl's interpretation of it, we take it that there are important lessons to derive from his model for current debates about explaining and understanding AI.[5]

Dealing with the artificial black box: explaining and understanding AI

Models of explaining and understanding developed for dealing with human psychopathology may provide a promising approach to address the challenges of black-box AI systems and can elucidate how human users can attempt to make sense of an AI's behaviour in two different, yet complementary ways. Going back to Jaspers' framework, we may take a new look at the problem of opacity. In accordance with Jaspers, we can distinguish two steps, the gathering of factual data and the establishment of relations between these data.

For the first step, we can distinguish between objective and subjective data. Objectively, we can observe the AI's *behaviour* by rigorous testing. Like with Jaspers, this objective stock taking should cover at least three different areas: (1) the AI's performance, measured e.g. by the accuracy of an AI's predictions, (2) its interaction with the world, measured e.g. by its behaviour in different settings, and (3), if applicable in instances such as the Deep Learning-based language model GPT3, the AI's *work*. On the subjective side based on phenomenology, our options for assembling factual data are necessarily limited: [6] We cannot grasp an ML models own perspective of their operation, unless we assume that the other mind is characterized to a large degree by human-likeness and has a similar capacity for consciousness (Shanahan, 2016). At least current AI models seem to lack both, barring us from a phenomenological *Vergegenwärtigung* of the machine mind. Here, Jaspers' model does therefore not offer any new insights.

However, we believe that Jaspers can contribute to a finer-grained analysis when it comes to the second step, aimed at establishing meaningful relations between factual data. It is here that we find room for Jaspers' distinction between understanding and explaining. The scope of explaining is in line with the many approaches of explainable AI that aim to establish general causal claims, in the sense of a "causality from without". Current approaches that e.g. use visualizations of weights given to specific factors to provide an "explanation interface" accessible to domain experts point in this direction (Holzinger et al., 2019). On an even more fundamental level, ML attempts to provide causal models by learning causal mechanisms would satisfy Jaspers' model here (Schölkopf et al., 2012; Parascandolo et al., 2018). However, as outlined above, causal explanations are only available to a limited extend in current machine learning practice, especially when it comes to deep learning.

Like in psychopathology, we should therefore embrace a two-pronged strategy to make sense of opaque machine learning models, based on both understanding and explaining, on causality from within and from without. In this sense, understanding should be conceptualized as a valuable complementary route to explainability, allowing us to identify meaningful, comprehensible relations, that may become immediately evident to us. An example by Jaspers himself may highlight how understanding can provide epistemic evidence. When examining the evidence of understanding, Jaspers refers to Nietzsche's use of genealogy, especially his Genealogy of Morality: "When Nietzsche's shows us convincingly how being aware of our own frailty, wretchedness, and suffering gives rise to about moral demands and religions […] we experience an immediate evidence that we cannot trace back any further." We understand the relation Nietzsche construes *evidently*.

Similarly, we may understand certain observable behaviours of machine learning models by examining its *genealogy* and its training history. Emily Denton and colleagues have recently suggested such an approach with view to the history of the ImageNet database (2021). Furthermore, if we engage in a form of intentional anthropomorphizing and follow the analogy of machine *learning*, we can also *understand* certain features by comparing the machine's learning to our own learning processes. For instance, we could infer from our own learning processes that an AI can only base its decisions and recommendations on its past experiences – similarly to training medical staff receives, improving their clinical decision making through experience over time: a diagnostic tool trained to distinguish photographs of (malign) melanoma and (benign) naevi may perform very badly in Black patients if trained exclusively on white patients – just like a human dermatologist who only received training using examples of lighter skin. Here, we understand the program intuitively, based on inferences informed by introspection, in a sense which Jaspers calls "causality from within". Mathematically, such understanding could also be fostered by what Angelov and colleagues call a "cardinally

[5] Jaspers' methodological convictions changed in the course of his life, and he moved away from his strict methodological dualism later in life (Schlimme et al., 2012). We still rely on this early model here since it seems most instructive with regard to medical ML models.

[6] To some extent, this is of course also true with view to the mind of other human beings, with the crucial difference that we are familiar with at least one human mind from an inward perspective: our own.

different approach to explainability" (2021): By choosing actual training data samples based on local peaks of the data distribution which they call "typicality", Angelov and Soares provide "prototypes" that are easily understandable by human users (2020).

Importantly, just like in psychopathology, such understanding may be empirically falsified (Ebmeier, 1987). Nietzsche's account of the genealogy of morality may be historically false in the particular instance of Christianity despite being understandable, as Jaspers notes (Jaspers 1946, p. 252). Similarly, looking at the genealogy of a training data set or prototypes among the training data could be misleading. It is therefore crucial to critically question the scope of understanding, as Jaspers repeatedly admonishes in critical remarks against Freud, and not jump to general causal rules. Also in machine learning, understanding demands to closely observe the program, its design and behaviour, or as Jaspers puts it: "understanding […] needs to be grounded in actual facts" (Jaspers, 1946, p. 255). [7]

Before we show how Jaspers' model can inform debates about understanding and explaining medical AI in particular, it seems imperative though to address the potential objection that we misguidedly anthropomorphise AI despite its non-human characteristics.

Understanding AI as misguided anthropomorphism?

There is an obvious caveat to discussing the relation of explaining and understanding AI with Jaspers. Jaspers originally discussed human psychopathology. We, however, want to draw on the relation of explaining and understanding with regard to AI. Indeed, the caveat is often brought up as a general objection to the use of human terms for artificial applications such as machine *learning* or artificial *intelligence.* These seem to be instances of anthropomorphism which is defined as "the attribution of distinctively human-like feelings, mental states, and behavioural characteristics to inanimate objects, animals, and in general to natural phenomena and supernatural entities" (Salles et al., 2020, p. 89).

The alleged threat of anthropomorphism to our adequate understanding of AI has been widely discussed (Salles et al., 2020; Watson, 2019; DeCamp & Tilburt, 2019) and anthropomorphism has been accused of being ontologically and morally dubious (Salles et al., 2020). The issue has been most prominently raised in relation to moral ascriptions, such as responsibility and trustworthiness, of algorithms. DeCamp & Tilburt (2019) have argued that this has severe consequences: "Trust properly understood involves human thoughts, motives, and actions that lie beyond technical, mechanical characteristics. To sacrifice these elements of trust corrupts our thinking and values" (p. 390). Similarly pointing out the differences between humans and algorithms, Watson (2019) writes: "Algorithms are not 'just like us' and the temptation to pretend they are can have profound ethical consequences" (p. 434). This finds expression in what Proudfoot (2011) calls the forensic problem of anthropomorphism, originally related to ascriptions of, say, intelligence to algorithms. As she writes:

"But how can a researcher's effort to 'convince himself or anyone else' of intelligence in machines be trusted if the researcher readily succumbs to anthropomorphism and make-believe—ascribing joy to a robot vacuum cleaner, for example?" (p. 952).

Generally, Proudfoot (2011) calls this the forensic problem of anthropomorphism which describes the risk of introducing cognitive biases in favour of the algorithm's intelligence by anthropomorphizing it. Unless the risk is mitigated, such judgements are deemed suspect. Is our attempt to understand AI similarly based on make-believe? After all, some may argue that it is an obvious mistake to discuss algorithms with regard to Jaspers' human psychopathology.

However, it is similarly dubious that the abolition of anthropomorphism is something that can be easily done. Proudfoot (2011) points out that even the critics of anthropomorphism in AI describe algorithms as stupid at the same time—a clear anthropomorphism as being stupid is a human characteristic. Our answer is that the employment of anthropomorphism should be pragmatic: if anthropomorphism is useful, it should not be jettisoned.

In the case of AI, there is some indication that it is. Bos et al. (2019) argue that anthropomorphism is an effective strategy for human participants to predict whether a *high-performing image classifier* AI model would label an image correctly. The participants of their study made reference to their own perception, either explicitly or implicitly, to predict the classifier's results. Interestingly, the researchers report that the mental model discussed "their own or general human abilities, indicating some cognitive separation of human and classifier abilities. The 'mental model' tag indicated awareness that participants were forming a mental model of the system as they did the task" (p. 954). This research is interesting for our context in at least two regards: first, it shows that anthropomorphism can be used for modelling in the context of AI, making use of what we, as humans, know about our own abilities. Anthropomorphism in this sense seems also in line with a current human-centric

[7] It is in this factual grounding that we can also situate the difference between understanding and interpreting: In absence of factual knowledge one may still provide a general interpretation ("deuten"), which lacks the properties of genuine understanding though (Jaspers, 1946, p. 252 f.; cf. Hoerl, 2013). The distinction between the two may not always be clear though, especially in the context of incomplete knowledge.

approach to explainability in AI "which treats it as a human-centric (anthropomorphic) phenomena rather than reducing it to statistics" (Angelov et al., 2021, p. 8).

Second, the study by Bos and colleagues also helps to disentangle the question of modelling from the question under which circumstances such anthropomorphist fiction constitutes an empirically effective strategy. After all, an intentional cognitive effort to understand AI by comparison to similar human abilities may not always be useful. Since anthropomorphist modelling is irrespective of the model's veracity, it will be important to distinguish between contexts in which accurate representation is required (Nguyen, 2020) while other models may benefit from "felicitous falsehoods" (Elgin, 2017). Bos et al. (2019) therefore rightly call for more empirical studies testing the factual effectiveness of anthropomorphist modelling in different contexts.

The human-centred distinction of explaining and understanding can therefore help to shed some light on explainability in AI. The discussion of understanding of AI should therefore not be hindered by general objections against anthropomorphism if it provides a useful tool. However, this still demands a clear conception of what explaining and understanding in relation to AI means.

Explaining and understanding medical AI

So far, we have sketched how a model developed by Jaspers in the context of human psychopathology can help to augment debates about explainable AI. Based on his distinction of explaining, aimed at general causal claims, and understanding, elicited by plausible evidence in singular cases, we advocate for methodological pluralism, harnessing both routes to establish meaningful relations between the factual data of machine learning. While we therefore started with a theory derived for a clinical purpose and employed it in the context of machine learning, we return to the clinic in this section, highlighting what Jaspers' model may imply for explaining and understanding medical AI. To do so, we draw on the well-known and widely cited example of IBM Watson for Oncology (WFO) and its shortfalls here (Strickland, 2019).

As we have seen, the first step of assessing such an AI will require careful observation of the program. These will contain different kinds of evaluations, both ex-ante and ex-post, to establish a factual basis for understanding and explaining. For instance, one would need to determine how the model and its hyperparameters were chosen, how it was optimized, and on which data, as much as one would need to evaluate its performance in different validation samples and identify the factors that had the largest impact on the model's prediction. To stay with the example, one would e.g. need to look closely at the health records which IBM used to train WFO, relying heavily on input from oncologists at the Memorial Sloan Kettering Cancer Center in the US (Jie et al., 2021), and at the model itself. To enable this kind of scrutiny, the program's developers would need to embrace open communication and share their "factual data" as openly as possible.

Having collected all this information, we would then have two routes to find meaningful relations in them. First, experts may aim for an *explanation* through an array of different methods (cf. Holzinger et al., 2019). Ideally, such an explanation would provide a general causal rule, which in turn may be used to improve the model. To stay with the example of WFO, it seems conceivable that by aiming for such a general causal rule, researchers may find a pattern in the program's decision that helps them to identify some novel (epi-)genetic causes underlying certain subtypes of cancer.

However, as a parallel, complementary approach, we should also aim at *understanding* the ML model. As we have shown at the beginning, to foster trust and enable important ethical goals such as informed consent, some grasp concerning the program's behaviour seems crucial for the end-users of a clinical ML application. As outlined, such an understanding can be based on plausible evidence, without establishing general causal claims – like we would, to use Jaspers' example, understand a connection between gloomy autumn weather and a tendency to commit suicide (Jaspers, 1946, p. 252 f.). In the case of WFO, such understanding may help us to make sense of observations that are immediately plausible to the lay person as well. A recent meta-analysis that compared WFO treatment recommendations with the recommendations of multidisciplinary teams of human experts found that concordance depended highly on regional differences and types of cancer. For instance, concordance of treatment recommendations was as low as 29.9% in gastric cancer, when comparing WFO with multidisciplinary teams from Asian countries (Jie et al., 2021). This observation becomes immediately plausible if one considers that WFO was trained and validated in the US and may therefore not agree with experts from other regions. After all, there are "large difference between the surgical methods and guidelines for adjuvant treatment of gastric cancer in China and the United States" (ibid.), and "WFO recommended the use of agents that are considered outdated in Korea" (Choi et al., 2019).

In such cases, we can understand the program's behaviour considering its training history, drawing on a form of "causality from within". Such understanding will require some form of knowledge about the AI model that can be related to our own reasoning processes, e.g. on which data it has been trained, where, by whom, and with which intentions. Other,

often more technical details may arguably not foster understanding, for instance, whether the underlying algorithm has been optimized using gradient descent, how many hidden layers were used in a deep learning architecture, or whether a sigmoid or a Rectified Linear Unit (ReLU) function has been used as activation function.

Like in psychopathology, it is important though to not mistake the evidence of understanding for the epistemic certainty granted by explaining (cf. Hoerl, 2013, p. 108). Jaspers notes this, when stressing that despite us understanding an autumnal death-wish, more people actually commit suicide in spring (Jaspers, 1946, p. 253). Similarly, we may also find that the underlying reason for WFO's problematic treatment recommendations in gastric cancers was not attributable to differences in regional treatment guidelines but based on the prevalence of particular mutations as has been reported for lung cancer (Jie et al., 2021).

Put differently, understanding does not imply giving up on causal explanations, just like for Jaspers understanding based on causality from within and explaining based on causality from without are not mutually exclusive. Yet, a complementary approach embracing both strategies to make sense of an AI model could prove fruitful in at least three ways. First, understanding meaningful correlations of an AI could be used to develop and test new hypotheses, thereby advancing genuinely causal explanations through the "encounter with the incomprehensible" (Jaspers, 1946, p. 254). Second, and particularly important in the context of medical AI, the differentiation between understanding and explaining could be seen as representing two different approaches tailored to different audiences. While explainability may continue to provide important technical tools for experts to improve and assess clinical AI, broader groups of end-users such as patients or physicians that do not command expertise in computer science may, at least partially, gain comprehension of an AI by means of understanding. Third, understanding and explaining could, in this sense, provide two complementary routes to increase an AI's trustworthiness: As recent research into the relation of explainability and trust has argued, the trustworthiness of an AI depends on both internal and external factors (Jacovi et al., 2021; Ferrario & Loi, 2021). While the internal trustworthiness of a model depends on the question whether the "reasoning process aligns with human reasoning" (Jacovi et al., 2021, p. 629) and may be promoted by a Jasperian understanding, the external path to trustworthiness relies on the observation from without and would therefore fall into the domain of what Jaspers calls explaining.

Jaspers' distinction between explaining and understanding, rooted in different accounts of causality, also connects well with recent philosophical contributions to the field such as Emily Sullivan's work on link uncertainty (Sullivan, 2020), despite terminological differences. As she convincingly argues, when discussing the black box problem of (medical) AI, one should distinguish between uncertainty introduced by a particular technical implementation – i.e., that we may not know how a particular deep learning model arrives at its predictions –, and link uncertainty, i.e. "the extent to which the model fails to be empirically supported and adequately linked to the target phenomena" (ibid.). Such link uncertainty can vary vastly in medical contexts. For instance, an opaque, deep learning-based program employed in pathology to diagnose cancer with rather clear aetiology and histological correlates acts on a fundamentally different link uncertainty than an algorithm employed in psychiatry to diagnose major depressive disorder. Given the many diverging levels of link uncertainty present in medical practice, is therefore crucial that the developers, users, and subjects of medical AI heed Jaspers' plead for methodological pluralism:

"All categories and methods have their specific purpose. It makes no sense to play them off against each other. Each of them has its own pure and appropriate realization, which is necessarily limited. Each of them, through absolutization, results in empty demands, ineffective talk and in modes of behaviour through which the free view of the facts is destroyed." (Jaspers, 1946, p. 384).

Conclusions

In this article, we have argued that the distinction between explaining and understanding as developed by Karl Jaspers in the context of psychopathology can provide a fruitful framework for current debates about the explainability of medical AI. In line with Jaspers, we have argued that explaining and understanding should be conceptualized as complementary epistemic approaches that must not be pitted against each other. We have shown how these approaches relate to current positions in the ongoing philosophical debate about medical AI and provided a practical example of its implications, drawing on IBM's Watson for Oncology as case study. Recent philosophical and ethical reflection on medical AI can therefore benefit from revisiting long-standing arguments from the philosophy of psychiatry to sketch a path towards ethically and legally sound, trustworthy AI in medicine.

Acknowledgements The authors would like to thank Bernice Elger, Eva De Clercq and two anonymous reviewers for their invaluable feedback on a previous draft of the paper.

Funding Open access funding provided by University of Basel. GS would also like to acknowledge funding from the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation's Grant Number: 32NE30_199436) to work on the revision of this paper.

Declarations

Conflict of interest None.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), 1–9

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, 130, 185–194

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424

Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J. E., Shaw, D. M., & Elger, B. S. (2022). Re-focusing explainability in medicine. *Digital Health*, 8, 20552076221074488.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford: Oxford University Press

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of medical ethics*. 2021;47:e3.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512

Bos, N., Glasgow, K., Gersh, J., Harbison, I., & Lyn Paul, C. (2019, November). Mental models of AI-based systems: User predictions and explanations of image classification results. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 184–188). Sage CA: Los Angeles, CA: SAGE Publications

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

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J. … Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8), 1301–1309

Choi, Y. I., Chung, J. W., Kim, K. O., Kwon, K. A., Kim, Y. J., Park, D. K. … Lee, U. (2019). Concordance rate between clinicians and watson for oncology among patients with advanced gastric cancer: early, real-world experience in Korea. *Canadian Journal of Gastroenterology and Hepatology.* 2019:8072928.

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine.The Lancet Digital Health, 1(8), e390

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2), 20539517211035955

Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., & Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, 27(4), 592–600

Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, 297, 103498

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5), 329–335

Elgin, C. Z. (2017). *True enough*. Cambridge, MA: MIT Press

Ebmeier, K. P. (1987). Explaining and understanding in psychopathology. *The British Journal of Psychiatry*, 151(6), 800–804

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(7639), 115–118

Ferrario, A., & Loi, M. (2021). The meaning of "Explainability fosters trust in AI". Available at SSRN 3916396

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V. … Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707

Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology*, 33(1), 1–3

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V. … Schafer, B. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707

Gough, J. (2021). On the proper epistemology of the mental in psychiatry: what's the point of understanding and explaining? *The British Journal for the Philosophy of Science* (accepted). doi: 10.1086.715106

Hoerl, C. (2013). Jaspers on explaining and understanding in psychiatry. In Stanghellini, G., & Fuchs, T. (Eds.). (2013). *One century of Karl Jaspers' general psychopathology*. Oxford: Oxford University Press.107–120

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine.Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312

Husserl, E. (2020). *Studien zur Struktur des Bewusstseins: Teilband III Wille und Handlung Texte aus dem Nachlass (1902–1934)*. Edited by U. Melle, & T. Vongehr. Cham: Springer

Hyland, S. L., Faltys, M., H ser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., Merz, T. M. (2020) Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine* 26(3) 364-373 10.1038/s41591-020-0789-4

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624–635)

Jaspers, K. (1946). *Allgemeine Psychopathologie* (4th ed.). Berlin: Springer

Jie, Z., Zhiying, Z., & Li, L. (2021). A meta-analysis of Watson for Oncology in clinical application. *Scientific reports*, 11(1), 1–13

Knoops, P. G., Papaioannou, A., Borghi, A., Breakey, R. W., Wilson, A. T., Jeelani, O. … Schievano, S. (2019). A machine learning framework for automated diagnosis and computer-assisted

planning in plastic and reconstructive surgery. *Scientific reports*, 9(1), 1–12

Kumazaki, T. (2013). The theoretical root of Karl Jaspers' General Psychopathology. Part 1: Reconsidering the influence of phenomenology and hermeneutics. *History of Psychiatry*, 24(2), 212–226

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8, 700

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, 49(1), 15–21

Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*, 279–288

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/>

Nguyen, J. (2020). Do fictions explain? *Synthese*, 199, 3219–3244

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29(3), 441–459

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2018, July). Learning independent causal mechanisms. *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, 4036–4044

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, 107, 101901

Proudfoot, D. (2011). Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5–6), 950–957

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, 11(2), 88–95

Schlimme, J. E., Paprotny, T., & Brückner, B. (2012). Karl Jaspers. *Der Nervenarzt*, 83(1), 84–91

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On causal and anticausal learning. *29th International Conference on Machine Learning (ICML 2012)*. , 1255–1262

Shanahan, M. (2016). Conscious exotica. *Aeon*.https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there (6.4.2021)

Spano, N. (2021). Volitional causality vs natural causality: reflections on their compatibility in Husserl's phenomenology of action. *Phenomenology and the Cognitive Sciences*, 1–19. doi: https://doi.org/10.1007/s11097-020-09724-9

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, 56(4), 24–31

Starke, G. (2021). The Emperor's New Clothes? Transparency and Trust in Machine Learning for Clinical Neuroscience. In: Friedrich, O., Wolkenstein, A., Bublitz, C., Jox, R.J., Racine, E. (eds.), *Clinical Neurotechnology meets Artificial Intelligence.* Advances in Neuroethics. Cham: Springer. 183–196.

Sullivan, E. (2020). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. doi: https://doi.org/10.1093/bjps/axz035

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44–56

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS medicine, 15(11), e1002689

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99

Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines*, 29(3), 417–440

Windelband, W. (1980). Rectorial Address, Strasbourg, 1894. Translation by Guy Oakes. *History and Theory*, 19(2), 169–185