Philosophy & Technology (2024) 37:105
https://doi.org/10.1007/s13347-024-00790-4

RESEARCH ARTICLE

# Crossing the Trust Gap in Medical AI: Building an Abductive Bridge for xAI

**Steven S. Gouveia**[1,2] · **Jaroslav Malík**[3]

Received: 18 February 2024 / Accepted: 27 July 2024 / Published online: 20 August 2024


## Abstract

In this paper, we argue that one way to approach what is known in the literature as the "Trust Gap" in Medical AI is to focus on explanations from an Explainable AI (xAI) perspective. Against the current framework on xAI – which does not offer a real solution – we argue for a pragmatist turn, one that focuses on understanding how we provide explanations in Traditional Medicine (TM), composed by human agents only. Following this, explanations have two specific relevant components: they are usually (i) social and (ii) abductive. Explanations, in this sense, ought to provide understanding by answering contrastive *why-questions*: "Why had P happened instead of Q?" (Miller in AI 267:1–38, 2019) (Sect. 1). In order to test the relevancy of this concept of explanation in medical xAI, we offer several reasons to argue that abductions are crucial for medical reasoning and provide a crucial tool to deal with trust gaps between human agents (Sect. 2). If abductions are relevant in TM, we can test the capability of Artificial Intelligence systems on this merit. Therefore, we provide an analysis of the capacity for social and abductive reasoning of different AI technologies. Accordingly, we posit that Large Language Models (LLMs) and transformer architectures exhibit a noteworthy potential for effective engagement in abductive reasoning. By leveraging the potential abductive capabilities of LLMs and transformers, we anticipate a paradigm shift in the integration of explanations within AI systems. This, in turn, has the potential to enhance the trustworthiness of AI-driven medical decisions, bridging the Trust Gap that has been a prominent challenge in the field of Medical AI (Sect. 3). This development holds the potential to not only improve the interpretability of AI-generated medical insights but also to guarantee that trust among practitioners, patients, and stakeholders in the healthcare domain is still present.

**Keywords** Medical AI · Abduction · Trust gap · Explainable AI · Pragmatism

Extended author information available on the last page of the article

## 1 Introduction

AI systems are becoming increasingly more capable in many fields and domains. Most recent evidence of this has been the rise of Large Language Models (LLMs; Zhao et al., 2023) like the GPT family of models (Radford et al., 2018; OpenAI, 2023; Bubeck et al., 2023). As these systems continue to improve, it becomes more important to consider how they are deployed. In this paper, we concentrate specifically on how AI should be used in medicine. Because of their greater image recognition and classification performance, several AI systems have already been used within the Medical Field (Pesapane et al., 2018; Briganti & Moine, 2020; Piccialli et al., 2021). However, current state-of-the-art AI, specifically neural networks, face a particular issue when deployed. The problem is that existing neural networks (NNs) are *black box* systems (Adadi & Berrada, 2018; Carabantes, 2020) that are opaque even to their designers. While modern AI systems are becoming more capable, the price for this capacity is the increasing complexity of these systems. As they continue to grow in size, we are less able to interpret the causes of their outputs. This problem gave rise to the field of Explainable Artificial Intelligence (xAI; Schwalbe & Finzel, 2023). It is clear that if we are going to deploy these systems in fields such as medicine, where human lives are at stake, it is necessary to develop methods of explaining these opaque AI systems.

However, what is an explanation? If *explainability* is our goal, we require a clear definition of it. One major problem with research in XAI is that it lacks an agreed-upon definition (Lipton, 2018; Mittelstadt et al., 2019; Miller, 2019). For this reason, we can observe a *pragmatic turn happening* in xAI (Paéz, 2019; Durán, 2021; Kim et al., 2022; Nyrup & Robinson, 2022), which aims to situate the notion of *explainability* in how people offer explanations on a day-to-day basis. It is argued that explanations have to be understood within a specific context and perspective. In this paper, we argue along similar lines. We take the stance that it is important to develop further xAI techniques on the background of the patient-doctor relationship involved in medical practice.

Medical doctors and healthcare professionals require the consent of their patients to do their work. To do so, they need to be able to *justify* and provide *reasons* for their diagnoses. In this way, the doctor-patient relationship is built on *trust* in the medical process and its ability to *justify* itself at every step. However, the patient-doctor relationship is jeopardised once NNs become involved. With the current opaque AI, no reasons for a decision can be easily extracted. Thus, *trust* is undermined in two specific ways: (i) the trust that patients can have in healthcare professionals in general and (ii) the trust that medical doctors can have in the medical process itself. This generates a Trust Gap (TG) that needs to be crossed when a patient asks for a reason for a specific procedure, treatment or diagnosis. To cross this gap, we must understand the process underlying common explanations.

In the first section of the paper, we explain the problems with the explanations offered within xAI. We argue that current xAI methods do little to help. The explanations on the offer do not reduce the ambiguity of AI decisions. Based on this, we must turn to social sciences and base our definition of *explainability* on everyday explanations. Following Miller (2019), we understand explanations as answers to contrastive

*why-questions.* We hold that the explanatory process comprises two distinct parts, social and abductive, that work in tandem to produce an understanding.

In the second section, we apply this framework to medical practice. We argue that there is a good case to be made that medical explanations are abductive, which, alongside their social character, allows doctors to cross the TG above. We make this argument in spite of the fact that abductive reasoning remains underexplored within the relevant literature, which partially also motivates our present approach. It is essential to differentiate between an Easy Trust Gap (eTG) and a Hard Trust Gap (hTG). An eTG is related to the fact that even if doctors can justify and answer *why-questions*, they can still face eTGs, since they may not be in control of the entire epistemic network of information and technical details of a specific medical case. However, because someone in the justificatory chain can grasp the necessary epistemic details that are missing, the TG can be quickly closed since the medical doctor can look for the missing information with another medical expert or technician. Within eTGs, the medical doctor has epistemic potentiality: a specific medical detail can be missed that can be filled by the epistemic role of another healthcare expert. The case is different when we consider hTGs: because AI models are epistemically opaque by nature (Carabantes, 2020), the epistemic potentiality is missing, and the missing information to provide an explanation to patients cannot be found anywhere in the epistemic chain.

In the third section, we apply these ideas to xAI by asking whether current AI systems (specifically transformer NN architecture) can transform those hTGs into eTGs. We hold that it is not enough to merely engage with *pragmatic turn* in xAI, but it is also vital to ascertain how much the current paradigm in AI is capable of following these pragmatic demands. If LLMs and transformer architecture generally show a promising capacity to engage in abductive reasoning, it would place transformers where they may contribute significantly to xAI. In this section, we do a literature review of recent studies testing the performance of transformers in abductive reasoning (Bhagavatula et al., 2020; Hessel et al., 2022; Liang et al., 2022) and social reasoning (Sap et al., 2022; van Duijn et al., 2023; Shapira et al., 2023) tasks.

Importantly, we find that while these systems can potentially use abductive reasoning, they also show some limitations. Transformers are capable of *selective* abduction because they can choose well from a set of pre-made hypotheses. However, they do not fare as well in *creative* abduction in tasks requiring the generation of novel hypotheses. We conclude that this deficiency is likely to have repercussions when these systems are tasked with generating explanations of their output. These limitations would likely affect the application of these systems in xAI. Therefore, we hold that there is a need for more research to improve current AI systems' creative abductive reasoning ability.

## 2 Explaining Away AI Ambiguity

In recent years, the field of xAI has received considerable critique for its approach. One of the identified issues is that the core concepts of the field, such as *explainability* or *interpretability*, show themselves to be "simultaneously important and slip-

pery" (Lipton, 2018, p. 20). These terms lack an agreed-upon definition, resulting in many xAI proposals that do not move the field forward. Miller et al. (2017) identify that the root cause is that the *inmates are running the asylum* in xAI, where explanations are formulated by data scientists for other data scientists. Many researchers working on xAI tend to take the approach of "I know it when I see it" when choosing their model of substantive explanation. However, that leads to researchers following their intuitions, with the result often being explanations unsuitable for the end users.

An excellent example of where current xAI methods fail is in cases where doctors have to interpret the outputs of NN. Often, doctors have to deal with *ambiguity* regarding the causes of the given output (Lebovitz, 2020). *Prima facie*, the information from AI should improve and speed up the diagnostic process. Nevertheless, this is only the case if the doctor arrives at the same conclusion as the AI. In cases where the AI disagrees with the clinician, it is necessary to interpret *why* the system generates the given output. This leads to a surge of *ambiguity*, where the doctor must doubt the system or her diagnosis. The problem is that the xAI methods available do little to diffuse this ambiguity; for example, in their survey, Wysocki et al. (2023) found that explanations provided by the CORONET system did not raise the understanding of the system itself. We believe that the problem lies in the form xAI explanations take.

The fact is that xAI methods concentrate primarily on identifying the importance of different input features. However, this has its issues, as noted by Lim et al. (2019, 4) in xAI "showing feature attribution or influence is very popular, but this only indicates which input feature of a model is important or whether it had positive or negative influence towards an outcome." We argue that the problem is that many xAI methods often, at best, answer *how-questions*, which "determine the set of causes that, if removed, would prevent the event from happening" (Miller, 2019, p. 7). A problem with this is that it generates an *interpretability gap* (cf. Holm, 2023; Ghassemi et al., 2021). Let us take saliency methods as an example. When the method generates a heatmap of important features in the input, the problem we encounter is that "it is not clear from the heat map *what* it is about the hot areas that the AI considers important for its output." (Holm, 2023, p. 3, emphasis original). We know what caused the AI to make a particular classification, and we gain knowledge of *how* AI processes the input, but we do not know *why* these features are important. This gap in our understanding is the source of the *ambiguity* that doctors face when interpreting AI, leaving it to them to interpret the feature's importance. That leads to one of the two results were either doctors understand these systems through their own *confirmation bias* (Ghassemi et al., 2021), where they think that what matters to them is also salient for the AI, or they let these systems decide for them, succumbing to the *automation bias* (Goddard et al., 2021; Lyell & Coiera, 2017) and trusting these systems unquestioningly.

This ambiguity is likely to affect the diagnostic process and the relationship between patients and doctors as a whole. We believe *the interpretability gap* constitutes the difference between an eTG and an hTG. To deal with the doubt generated by AI, we must base our understanding of explanation on "what it means to say that a human agent *understands* a decision or a model" (Paéz, 2019, 442, emphasis original). For this reason, several researchers call for a *pragmatic turn* in xAI (Paéz, 2019; Durán, 2021; Kim et al., 2022; Nyrup & Robinson, 2022). They argue that it is nec-

essary to develop xAI from the perspective of the explanatory processes that human beings use and base on it the account of human-computer interaction. Further, it is argued that we need to engage with models of explanation in social sciences (Miller et al., 2017; Miller, 2019). When we do, we find that there are four criteria for explanations: (1) Explanations are *contrastive*; (2) Explanations are *selected*; humans are biased toward particular *causes*; (3) Probabilities and statistical data do not generate understanding; and (4) Explanations are *social*.

In this paper, we are particularly interested in (1) and (4) since we believe both are important for dealing with the *ambiguity* we encounter with Medical AI and deciding whether we are facing an eTG or hTG. We must understand that an explanation is both a process and a product (Lombrozo, 2006). The product of an explanation takes the form of an answer to a *why-question*, which is *contrastive* in the sense that people do not want to "explain the causes for an event per se, but explain the cause of an event *relative to some other event* that did not occur" (Miller, 2019, p. 9, emphasis original). Explanation ought to answer the question of 'Why had P happened instead of Q?' (cf. Lipton, 1990; Miller, 2021). We can observe that this is the case with *ambiguity* with medical AI. The clinicians in those situations seek to understand why AI came to a different conclusion than they did. They seek an explanation on the background of their own diagnosis and want to know why AI thinks differently. That creates the context for the given explanation. Current xAI methods do not answer this question, thus generating the issues we discussed above.

As a process, the explanation comprises two different but connected explanatory processes. Explanation involves "transferring knowledge between *explainer* and *explainee*, generally an interaction between a group of people" (Miller, 2019, p. 6, emphasis added). Thus, the explanatory process is social; explanations "are communicative acts where an *explainer* conveys some information to an *audience*" (Nyrop & Robinson, 2022, 13; emphasis original). Following this, within the *contrastive* context, explanations must also be aimed at a particular *explainee*. This, however, has to go beyond merely producing an explanation and be understood as an ongoing process. As urged by Rohlfing et al. (2021), we simply cannot view the *explainee* as a passive receiver of explanation; instead, she and the *explainer* steer the process of explanation together. Both of the participants in the explanatory process are social agents. Therefore, the *explainer* needs to model the *explainee* and vice versa to converge on the *explainee* gaining understanding.

The part that we want to stress is that *explanations* happen within social practices,[1] which involve "a protocol that serves as an orientation in the sense of what to do next and an interpretation of where an action could be going" (Rohlfing et al., 2021, 724). For human beings, these social practices very often have a linguistic form. Linguistic social practices help to further structure and control the process of explain-

[1] Several discussions in this paper highlight the active role of the explainee in explanations, emphasizing that explanations are not mere passive information but rather an ongoing dialogue between parties, which indicate the social nature of explanations that will be central to our argument. Kästner et al. (2021) introduced the Explainability-Trust Hypothesis (ET), positing that explanations can foster trust in the explainee. Blanco (2022, pp. 249–251) further examined this idea, emphasizing the importance of considering the explainee's role and background knowledge for ET to be effective. Both theses can serve as a background assumption for the argument we are developing in this paper.

ing. Because of this, developing xAI by engaging with how natural language structures explanations is likely to prove fruitful. Several studies have shown that natural language explanations have benefits, such as increasing trust in the given system (Chaves & Gerosa, 2020; De Gennaro et al., 2020). However, current xAI primarily uses natural language for "generating explanations using context and presenting them as a *fixed* explanation" (Cambria et al., 2023, p. 8, emphasis added). This is problematic because *explanations* are primarily *conversations*. As shown by Lebovitz (2020), when doctors interact with AI, they are often frustrated precisely because they lack any means of having a recourse with the AI. Only presenting the explanation in natural language is not enough to deal with the *interpretability gap*. We need xAI systems that are conversational agents (Nguyen et al., 2023; Jentzsch et al., 2019) capable of engaging in the social practice of explaining. We would add that *explainable* AI may not be enough; we also need *explanatory* AI (Elton, 2020; Sovrano & Vitali, 2022; Grangea et al., 2022), where we have autonomous AI agents which are capable of explaining other artificial agents or themselves.

However, while the social process can help the *explainer* and *explanee* eventually arrive at a satisfying explanation, it cannot be the start of the explanatory process. The social process allows for knowledge transfer, but it cannot fully generate the explanation. The second explanatory process is a cognitive process, which takes the form of "abductive inference for 'filling the gaps' to determine an explanation for a given event" (Miller, 2019, p. 6; cf. Lombrozo, 2012). Abduction involves the agent encountering an unexpected outcome, which prompts her to develop a possible hypothesis of its cause. According to Hoffmann et al. (2020) and Miller (2023), the process of *abduction* has the following five phases:

1. *Observation of an event* – The agent encounters an event that is not in accordance with its current mental model, which prompts the agent to start an exploratory process, where she imagines the causes of said event.
2. *Generate Hypotheses* – The agent generates several informed guesses that could explain the event, which then serve as hypotheses moving forward.
3. *Judge the plausibility of the hypotheses* – The agent starts looking for evidence that could prove or disprove its formulated hypotheses, with some of them discarded as other hypotheses are found more probable.
4. *Resolve understanding* – In the end, the agent finds one hypothesis that explains the event particularly well and fits the evidence found. This hypothesis is then adopted as the explanation of the event. The abductive process may stop here or proceed to 5.
5. *Extend* – Revise and extend the process as new evidence is encountered by the agent.

What is argued by several researchers is that an *abductive* process of this form may serve as a good foundation for xAI (Miller, 2019, 2023; Hoffmann et al., 2020; Medianovskyi & Pietarinen, 2022). In xAI, the hope is that if we produce enough local explanations for specific AI outputs, we could eventually arrive at a general model of the given system. However, that is not the case, as pointed out by Mittelstadt et al. (2019); these local explanations have issues with generalisation precisely because

they are not *contrastive*. A particular set of data points may lead us to one conclusion about the importance of an input feature, but we need to ask ourselves the question of "[importance] compared to what?" (Mittelstadt et al., 2019, p. 282). We need to situate such inductive explanations in some greater context. That is why, if we want to develop xAI, what is necessary "is not induction but abduction" (Medianovskyi & Pietarinen, 2022, p. 3). The reason is that abduction serves as the beginning of the explanatory process because "the conclusion we arrive at is *a* theory, not *the* theory. It is *a* case, but not the *only* case. Abduction provides possible accounts for *why* something might be the case." (Veen, 2021, 1176, emphasis original). Abduction gives us a possible explanation, which we are invited to investigate. That further frames the explanatory context. However, we should not think that abduction captures the whole process of generating knowledge:

"Abductions alone do not add to present state knowledge, the full cycle of abduction-deduction-induction does. Abduction concludes with a question and is an invitation to further inquiry. Its conclusion proposes, could X possibly be the case? Let us find out." (Medianovskyi & Pietarinen, 2022, p. 7).

Abduction alone cannot generate knowledge, but it is an important starting point. Interestingly, Medianovskyi and Pietarinen (2022) argue that abduction, combined with current xAI methods such as feature visualisation, could help facilitate the finding of explanatory hypotheses. Current xAI methods may help us generate observations, which can further motivate specific hypotheses regarding AI behaviour. While insufficient on their own to provide illuminating explanations, they can provide support for abduction. In xAI, abduction may be vital because it can supply a possible *why* and kickstart the exploratory process that characterises explanations.

If Medianovskyi and Pietarinen (2022) are correct, it is important to study and eventually develop abductive reasoning processes within current and future AI systems that would supplement and scaffold the xAi methods already used today. Further, following the argumentation of Miller (2019), the presence and active use of abductive reasoning in AI and xAI is likely to prove useful for translating the decisions of AI systems into an explanatory form that is more accessible to the end users of these systems. In this manner, abduction may augment current approaches in significant ways that could alleviate many of the difficulties mentioned above.[2] We interpret the *pragmatic turn* and its concept of explanation (so far as it is inspired by findings in social sciences) as putting forward the claim that social and abductive processes of explanation ought to be taken seriously. While current xAI methods do provide some limited insight, their explanatory value (and the trust that is generated as a result of successful explanation) can be improved if we account for the explanatory processes in use.

[2] As pointed out by one of the anonymous reviewers, focusing on abductive reasoning and explanatory process also has one particular advantage over a more standard approach. In xAI, a causal explanation is usually pursued as the desired result of the xAI methods. It makes sense to desire to have an account of what particular cause led to the event. However, *causal* explanation is very demanding, especially in the face of incomplete evidence (which is usually the case in healthcare contexts). So often, the best we can do is to guess (cf. Popper, 1959, p. 278), which human reasoners often do. By focusing on the use of abduction in xAI, we relate AI more closely to human practices in medicine.

However, the particular way the social and abductive processes are going to be implemented in xAI is likely to produce important caveats that we need to pay attention to. A potential issue with supplemental xAI methods added on top of an AI model is that they only produce an *interpretable representation* of the system's features. The fact of the matter is that "**interpretable** explanations need to use a representation that is understandable to humans, regardless of the actual features used by the model" (Ribeiro et al., 2016, p. 3; emphasis original). Whatever insight that is produced by external xAI systems is not direct in the same way as human abductive reasoning. Human reasoners often reflect on their own cognitive grasp of the problem at hand in ways that are unlike how the AI model is interpreted by the xAI system. This significantly complicates the picture of human-computer interaction. While two human interlocutors can hold each other accountable directly, with AI, one can only interact through layers of *interpretable representations*. Therefore, we would argue that xAI methods cannot be just bolted on already existing systems. Rather, following our appeal to the concept of *explanatory* AI, what we need to ultimately aim for are AI systems that, in themselves, are responsive to the accountability necessary for trust.[3]

So, following all of this, we understand explanations as involving *distinct social and abductive processes which aim at providing answers to contrastive questions*. Our task is now two-fold. First, it is necessary to judge how the practice of explanation in medicine reflects the criteria set out above. In line with the *pragmatic turn*, we have to pay attention to the particular context of explanations we find in medicine. We, in particular, set our sights on the doctor-patient relationship because we believe it may reflect some issues similar to those we encounter in xAI. Building on how doctors can explain their hypotheses to patients may provide a valuable blueprint for xAI. Second, we need to judge how capable the current AI is of following this standard and how its capability could be applied to medical xAI. That is done in the third section. However, first, we must show why abduction is a relevant tool for medical sciences.

[3] Accountability relates to one philosophical issue connected to the TG. Our claim in this paper is that the hTG presented by current *opaque* systems can be turned into eTG if the xAI system engages in abductive reasoning to generate a *why-explanation*, good enough to serve as an explanatory reason for AI output. The problem is that if xAI systems only produce representations, we cannot say that the hTG has been resolved into eTG because there is still "no one" who knows *why* precisely the AI system produced a particular output. It could be argued that, at best, we created a system that "mimics" human abductive reasoning and provides probable explanations. This is a complex issue related to the overall *opacity* of AI systems. While solving all of this together is outside the scope of this paper, we would like to stress that we do not make only an appeal to abductive aspects of explaining but the social ones as well. In order to deal with the TG, xAI developers also have to understand that their models ought to be bound to social practices. While abductions generate explanations of the AI models, these explanations are subject to social norms and standards. Medicine is no different. The framework within this paper cannot solve the entire problem of xAI. However, that is not our goal. Rather, we aim to formulate a set of standards for xAI that would be sufficient for their ethical deployment in healthcare, and this can be considered an improvement over the current state of affairs.

## 3 Abduction in Medical Sciences

There has been a long and controversial debate about the nature of medical reasoning in medical sciences since it is a concept investigated by different disciplines with different aims (cognitive psychology, medical education, philosophy of science, etc.) (Norman, 2005). The nature of medical reasoning involves – most of the time – a complex interplay between deductive, inductive, and abductive reasoning in a heterogeneous cycle. Each kind of reasoning has an essential role in diagnostic and decision-making processes within the field of medicine:

(i) Deductive Reasoning: Deduction works by applying general principles to specific cases, following a top-down approach; for example, if a patient exhibits specifically known symptoms (S ∈ x, y, z) and we know that (S ∈ x, y, z) are associated with a particular condition A, then, by the deduction, the medical doctor concludes that the patient has that medical condition A;
(ii) Inductive Reasoning: Induction works from drawing a general hypothesis from specific cases and observations, following a bottom-up approach; for example, if a physician observes several patients with the same (S ∈ x, y, z) that are responding well to a treatment B, then it can be induced that the medical treatment B is effective for that medical condition A;
(iii) Abductive Reasoning: Abduction works by seeking the most plausible explanation from an incomplete set of evidence or medical observations; for example, given an incomplete set (S ∈ x, y, z, … *n*), the physician can provide a potential diagnosis that best explains (S ∈ x, y, z, … *n*), leading to further confirmation mechanism for that specific diagnosis.

In essence, medical reasoning is a complex process that integrates deductive reasoning to apply general medical knowledge, inductive reasoning to form hypotheses based on specific cases, and abductive reasoning to generate the best explanations in the presence of medical uncertainty. Undoubtedly, each of these reasoning processes plays an important and relevant role within the medical enterprise. In medical diagnostics, however, following Gungov (2018), combining deductive and inductive methods is considered insufficient for ensuring favourable medical outcomes. For our argument, we need to provide further reasons and evidence that, at least in some parts of the field of medicine, abduction plays an essential and fundamental role in providing a foundation for generating explanations. We have to do so in spite of the unfortunate fact that "clear and explicit analyses of abduction in clinical research and practice are not very frequent" (Chiffi & Andreoletti, 2023, p. 416).

Abductions[4] are known to be an interesting cognitive instrument for healthcare professionals since they allow further exploration when medical uncertainty is present in many medical judgments concerning prognosis, treatments or diagnoses (Chiffi,

[4] Ramoni et al. (1992) make a distinction between unfocused and focused abduction: the first is related to the selection of general hypotheses to constrain the range of possible alternatives in the first stage of the medical process; the second is related to the idea that those selected general hypotheses need to be shaped on the clinical characteristic of the patients (Ramoni et al., 1992). Although this distinction is pertinent, it is not relevant to the general argument of the paper.

2021). Abduction can also help clinicians in choosing relevant aspects to understand the interplay between evidence collected at the population level and the specifics of a single clinical case – that can be quite dissimilar from each other – crucial for formulating judgments on individual risk (Chiffi & Andreoletti, 2023).

It is essential to notice that we are not arguing that all medicine is abductive but that a crucial part is indeed based on abductive reasoning in conjunction with induction and deduction. While induction derives a general truth from facts and evidence, and deduction explains something from a general rule, abduction creates an additional layer that is missing by interpreting a specific case about a higher-level hypothetical pattern that can be confirmed or rejected with further mechanisms (e.g. testing, exams, evidence) (Eriksson & Lindström, 1997).

Following Karlsen et al. (2020), we can provide a specific medical diagnosis example that illustrates the relationships between the three kinds of reasoning and their conjunctions in cases of medical uncertainty. The process can be divided into four specific stages:

Stage 1: Specific symptoms are considered incomplete or unexpected, requiring a new explanation that diverges from an established hypothesis.
Stage 2: A new hypothesis, strategically formulated to predict and elucidate the incomplete set of symptoms, is postulated through the process of abduction, allowing for a comprehensive exploration of potential explanations.
Stage 3: This procedural sequence advances through deduction, navigating toward the identification of the hypothesis's most probable explanation.
Stage 4: When tests consistently validate the predictions, the outcomes are incorporated into scientific findings through an inductive process within a diagnosis, effectively elucidating the symptoms in the most comprehensive manner. (cf. Karlsen et al., 2020)

Charles Peirce, the pragmatist philosopher, was one of the first authors to define abduction, claiming that it is "the operation of adopting an explanatory hypothesis", providing a first formal description of it: "The surprising fact, C, is observed; But if A were true, C would be a matter of course, Hence, there is reason to suspect that A is true" (Peirce, CP 5.189).[5] This definition fits exactly the diagnostic example provided above and many of the descriptions that can be found in the medical literature about the medical process of reasoning: for example, Aliseda (2006) claims that "abduction is a reasoning process invoked to explain a puzzling observation" (Aliseda, 2006, p. 28).

Another important characteristic of abduction[6], following Peirce (cf. CP 6.525), is that it has a selective component that is relevant to making it a valuable instrument

[5] Peirce's definition of abduction changed over time, as shown by Pietarinen and Bellucci (2014), but this change does not affect our general claim about the role of abduction in medical sciences.

[6] There is a long debate about whether abduction is similar or different from Inferences to the Best Explanation (IBE). Again, this conceptual debate is intriguing but not primarily relevant to the paper's central argument. Douven (2021) argues that they are similar; Campos (2009) claims they are different. A more plausible position between these two extremes is to argue that abduction and IBE overlap somewhat, even if they are also different (Mackonis, 2013, p. 976).

in a medical context. Since an unknown medical condition can raise many different hypotheses to explain that medical condition, abduction needs to not only generate those hypotheses but also provide a criterion to rank or select the most promising hypotheses that should have medical priority over the others (Stanley & Nyrup, 2020), in this sense, abduction is also a process of eliminating explanations that are less plausible until the most convenient one is achieved. Moreover, abduction can also be seen as *selective* in opposition to *creative*: in the first case, the medical doctor focuses on a specific hypothesis from a closed set of possible explanations that are already selected by previous cases or experiences; in the second case, the medical doctor infers a hypothesis which is not available at all (Magnani, 2001, p. 75).

Abductive reasoning should not be confined solely to medical diagnosis; it is also pertinent to other crucial medical tasks, such as implementing and adopting specific therapeutic strategies for a given disease. Take, for example, the real case of the Covid-19 pandemic (Ali et al., 2020). The standard therapeutic interventions proved ineffective in addressing this particular disease. There was limited evidence regarding its nature, a lack of hypotheses to explain high individual variations, and many traditional treatments for diseases with analogous symptoms were inefficient. In a nutshell, Covid-19 created a medicinal context of high uncertainty, and healthcare researchers had to employ abduction to generate new and alternative hypotheses to address this medical uncertainty. This involved selecting the most promising hypotheses given the available information constraints (Campaner & Sterpetti, 2023, p. 457).

Abduction not only plays a role for doctors in their pursuit of a correct diagnosis or therapy but is also crucial for the communication acts between doctors and patients: abduction is also used by patients so they can understand their medical situation. Based on real-case examples[7] reported by Martini (2023), we will show how abductive reasoning is used (i) in an actual diagnosis situation and (ii) in the understanding of a patient's own medical condition by themselves.

Regarding (i), a patient informs their doctor about their medical history, starting with back pain and leading to odd-coloured stools and dark urine. The doctor claims this progression was expected due to cancer's natural progression that tends to block the bile duct, connecting the abnormal test results (gamma-glutamyl transferase, phosphatase, bilirubin) to pancreatic cancer using abductive reasoning since it explains the abnormal results and also the symptoms described by the patient. This best hypothesis provides a unique diagnosis, predicts what might happen, and suggests treatment. It also helps the patient make sense of their own experience. A biopsy later confirms the initial best guess of pancreatic cancer, making the diagnosis even more certain (cf. Martini, 2023, 473–474).

Regarding (ii), a patient's partner informs the oncologist that the patient lost significant weight, attributing it to health issues affecting their mental health via anxiety. Despite anxiety being more common than pancreatic cancer (the correct diagnosis), it served as an explanation for the partner's abductive inferences about weight loss. Another example can be offered with a patient who, because of the COVID-19 situ-

[7] The cases are related to the COMMUNI. CARE protocol by recording and transcribing doctor-patient interviews at San Raffaele University Hospital (Consolandi et al., 2020).

ation, slept on the couch for about 20 days and experienced lower back pain. They attribute the pain to couch-sleeping, a more common explanation than pancreatic cancer, the correct diagnosis. In both cases, the most readily available explanation is used to support abductive inferences by the patients to explain their symptoms before seeing a healthcare specialist (cf. Martini, 2023, 476–477).

Finally, abductive reasoning is also relevant for clinicians when a patient's symptoms are not only explained by particular physiological factors: this allows healthcare experts to explore social dimensions influencing a patient's health. Some diseases can have multifaceted presentations, and since explanations are not isolated acts but are embedded in social interactions (Rohlfing et al., 2021), the best hypothesis to explain specific symptomatology can be based on a broad range of social influences, such as environmental factors or lifestyle choices of the patient.

This second section provides diverse reasons to consider that abductive reasoning is, in fact, one of the fundamental instruments for different kinds of healthcare practices in human medicine. In the next section, we provide some evidence that at least some AI systems can perform abductions. If that is the case, these technologies could be a valuable instrument to transform hTGs currently present in Medical AI into eTGs, making the process more transparent to medical doctors and, consequently, patients. However, the abductive reasoning of AI is subject to specific limitations with possible consequences to their use in xAI. We go into detail in the next section.

## 4 Abduction in Artificial Intelligence

In the previous section, we argued that there are good reasons to suppose abductive reasoning plays a significant role in medical practice, especially in communication between doctors and patients. Further, abductions were found to help reduce uncertainty in medical research and decision-making. If abductive inferences of the kind we examined are useful in this way, then it is necessary to judge how capable current AI systems are compared to human performance. In this third and final section, we discuss this possibility. To do so, engaging with the current architectural paradigm of AI is necessary. In the case that current AI displays these capabilities, they would play a relevant role in medical contexts.

The success of LLMs, such as various models of GPT (Radford et al., 2018; OpenAI, 2023; Bubeck et al., 2023), has been facilitated by their transformer architecture (Vaswani et al., 2017; Lin et al., 2021). Initially designed for machine translation,[8] transformers have proven very capable in natural language processing (NLP). The source of this success was the use of the novel technique of self-attention, which enables transformers to consider different parts of the input and weigh their importance for the output. Architecturally, this ability has been further extended through

[8] The vanilla transformer architecture was designed for sequence-to-sequence modelling and was comprised of encoder-decoder blocks. Since then, alternative transformer architectures have been designed for different NLP tasks. These are encoder-only transformers such as BERT (Devlin et al., 2019), used for natural language understanding and decoder-only transformers such as the GPT family (Radford et al., 2018), utilised for sequence generation. However, differences between these architectures are not crucial for the goal of this section.

multiple attention heads, which allow transformers to consider relationships of parts of the input from different perspectives. Transformers have since proven to be not only valuable for NLP but also in other areas, such as computer vision (Dosovitskiy et al., 2020; Khan et al., 2022) because they can accept different modalities of input as tokens (Xu et al., 2023). For our argument, we must examine whether transformers demonstrate promise along the dimensions we have stated before. If they do so, it will mean that transformers could be further exploited for better xAI techniques in healthcare. In this section, we thus review the recent literature surrounding the performance of transformers at social and abductive reasoning. While some of these studies are preliminary, they give us the most up-to-date picture of AI's abilities.

Let us start with the social aspects of explanations. Since human explanations often have a linguistic form, it is clear that LLMs could be employed to create xAI systems that generate explanations in the natural language form. LLMs could be well-suited to become *conversational agents* (Nguyen et al., 2023). However, their performance on the explanatory tasks ought to be questioned. For example, while Bubeck et al. (2023) show that GPT-4, in particular, was capable of seemingly answering *contrastive* questions of the kind we have in mind, they note that "GPT-4 often generates explanations that contradict its own outputs for different inputs in similar contexts." (Bubeck et al., 2023, p. 62). That means that these explanations are not *process-consistent* in the sense that they do not offer any insight into how the LLM arrives at its output, showing that, despite their linguistic form, explanations provided by LLMs do not fulfil the necessary conditions of being genuine explanations.

Further, we stressed that the social process of explaining requires mutual modelling between the *explainer* and the *explainee*: with LLMs, there is a debate about whether they are capable of such modelling. It has been argued by Shanahan et al. (2023) that we ought to metaphorically understand LLMs as playing the role of an agent required by the user's prompt. However, the metaphor invites the question of whether the linguistic capacity of LLMs is predicated on having particular models of linguistic agents (cf. Andreas, 2022). If so, it is necessary to judge how good LLMs are at this modelling; whether LLMs possess a Theory of Mind (ToM) may serve as a good proxy for this assessment. Several studies found smaller LLMs generally incapable of social reasoning requiring ToM (Sap et al., 2022; van Duijn et al., 2023; Kosinski, 2023). Larger systems like GPT-4 were discovered to be quite capable at some ToM tasks[9] with a performance at or above children's level of reasoning (Kosinski, 2023; van Duijn et al., 2023). What is problematic is that this performance significantly suffers when models are presented with adversarial examples (Ullman, 2023; van Duijn et al., 2023; Shapira et al., 2023). That suggests LLMs do not have a general ability to engage in social reasoning. However, as van Duijn et al. (2023) noted, we are still in the early stages of testing the social reasoning of LLMs: the fact that they perform well at some ToM tasks should not be discounted. LLMs continue

[9] These were the *Unexpected Contents Task (aka Smarties task)* and *Unexpected Transfer Task (aka the "Maxi task" or "Sally–Anne" test)* in Kosinski (2023). Ullman (2023) created alternative adversarial versions of these same tests. Van Duijn et al. (2023) similarly tested LLMs on the *Sally–Anne test* alongside the *Strange Stories test* and the *Imposing Memory test.* Sap et al. (2022) and Shapira et al. (2023) instead used datasets for AI (such as Triangle COPA, SocialIQa, ToMi, epistemic_reasoning, FauxPas-EAI) that were either explicitly developed by them or someone else.

to undergo rapid development; future generations of these systems could be more capable.

Further, there are possible avenues we could take to improve the ToM capabilities of LLMs.[10] One thing that van Duijn et al. (2023) suggest is that LLMs developed capacity for ToM during the fine-tuning, specifically during the RLHF (reinforcement learning from human feedback) phase, where the models were incentivised to generate outputs in cooperation with the intentions of human evaluators. If proven true, LLMs could be responsive to cooperative co-modelling that characterises explanations. Further literature and study could suggest various ways to capitalise on this potential. For example, Buckner (2024, 305–344) explores two distinct approaches in psychology and their implementation in AI. A more traditional "rationalist" approach supposes that ToM arises from a distinct need for social competition that forces agents to exploit their innate capability to distinguish themselves to model and predict other agents. In AI, we could take this approach to conceive of ToM as a problem of *inverse planning*, where the AI agent models other agents by deriving a *reward function* that would best predict other agents' behaviour; specifically, which function would provide the most reward for the displayed behaviour. Such a system takes the behaviour of other agents as input and produces a hypothesis of what motivates their behaviour as an output. An example of a system inspired by this approach is ToMnet from Rabinowitz et al. (2018). The drawback of this method is the need to develop additional system modules, which enable the given system to emulate innate human capability for ToM.[11]

The other approach that Buckner (2024) classifies as "empiricist" takes a more developmentalist perspective. Proponents of this view hold that human infants lack the means necessary to think of themselves as separate beings due to being dependent on a caregiver. Instead, social cognition arises due to the influence of affective states (such as sympathy) early in life, which help to form and shape the child's social cognition over time. In AI, this approach materialises as the field of *affective computing* (Picard, 2000), where we aim to enable AI to recognise the same kind of cues that allow humans to be sensitive to the mental states of other agents. An intriguing possibility that is suggested by Buckner (2024) is that the current RLHF method could be further improved and extended if we "find ways to leverage a wider range of qualitatively distinct valences of human feedback" (Buckner, 2024, p. 341) like these *affective states*—either of these aforementioned approaches present possibilities for improving upon ToM capabilities of current systems. Future research will have to decide which of these approaches is better.

[10] We would like to thank the reviewers for highlighting this important development, which should be considered in this debate.

[11] Rabinowitz et al. (2018) specifically discuss two additional modules on top of the basic *prediction net*: the *character state net* and the *mental state net.* The ToMnet is tasked with modelling agents moving within virtual grid worlds. C*haracter state net* serves to characterise the agents by watching the trajectories of their movements and creating an appropriate character embedding, which is then fed to the *mental state net*. This second component uses c*haracter embeddings* and current episode trajectory to formulate a *mental state embedding.* The *prediction net* utilises these *mental state embeddings* to predict future agent behaviour while the *embeddings* constantly update. Rabinowitz et al. (2018) hope this approach will approximate how humans mentalise ToM.

But what about the abductive reasoning capabilities of these systems? An important thing to consider is that the field of medical imaging – particularly radiology – saw the most widespread use and adoption of AI systems in medicine (Pesapane et al., 2018). Transformers have been shown to outperform the previous architectural paradigm used in this domain (Dai et al., 2021). Moreover, when transformers are considered in the medical field, it is for their capacity to accept different modalities of data (He et al., 2023). This quality is attractive in healthcare, where multi-modal data is increasingly more available (Acosta et al., 2022). Therefore, we must judge transformers' abductive reasoning capability in different modalities. The abductive reasoning capacity of transformers has been most explored in NLP (Bhagavatula et al., 2020). However, there are also prominent studies in visual abductive reasoning (Hessel et al., 2022; Liang et al., 2022). Below, we go over these studies that compare the performance of transformers to human abductive reasoning.

In NLP, we have an ART benchmark developed by Bhagavatula et al. (2020) to test the abductive reasoning of LLMs based on narrative text. The system is given a set of observations and is tasked with providing a plausible hypothesis which explains them. Bhagavatula et al. (2020) created two different datasets for this benchmark. The first one is the Abductive Natural Language Inference (αNLI) task, where a set of observations and two plausible hypotheses are presented to AI, and the system has to choose the more plausible hypothesis. According to the public leaderboard (Allen Institute of AI, n.d.), AI has surpassed the human performance of 92.90% on this task with a score of 93.65%. The problem is that the methodology of this result has not been published. However, we can find the methodology of the second-best result of 92.80% (Zhang et al., 2022). Even if we discard the unpublished best result, the performance of transformers proves incredibly close to human capability. This result must be contrasted with the second dataset, Abductive Natural Language Generation (αNLG), where no set of plausible hypotheses is provided to the system tasked with generating novel plausible hypotheses. Recent tests on this dataset (Zheng, 2023) show that transformers have reached a human approval of 79% but have not crossed the threshold of human performance of 96% set by Bhagavatula et al. (2020).

When it comes to testing visual abductive reasoning, we find two prominent datasets. Hessel et al. (2022) developed the SHERLOCK dataset to test the visual abductive reasoning of transformers. It consists of a large set of images, annotated with highlighting bounding boxes and includes textual clues that describe the highlighted regions. The dataset also includes a collection of commonsense abductive inferences prompted by the boxes and clues.[12] This dataset tests AI on three tasks, but we are only interested in two tasks that compare AI and human performance. In the *localisation task*, AI has to align the correct inferences with the boxes that best support them. AI is either given the bounding boxes in the correct regions or has to choose between different randomised proposals of how the boxes could be arranged. In the latter case, the AI is also tested on how closely its chosen proposal intersects with the position of the original boxes. On the other hand, the *comparison task* involves AI judging the plausibility of a set of hypotheses for the image. This judgement is then compared to

[12] All of the highlights, textual clues, and abductive inferences were provided by human annotators during the creation of the dataset.

assessments made by human evaluators via pairwise accuracy, where 50% is subsequently subtracted to see how much better the AI model performs over mere chance. In *localisation*, the best human score was 92.30% with boxes placed and 96%[13] without them. Meanwhile, human agreement arrived at 42% (92%) in the *comparison task*. Currently, the best overall AI performance (Allen Institute of AI, n.d.; Zhang et al., 2024) is the accuracy of 89% (boxes placed) and 44% (without placed boxes) in the *localisation* task, while the *comparison* accuracy score is 31% (81%).

Finally, the other dataset is VAR (visual abductive reasoning), created by Liang et al. (2022). VAR is comprised of video examples that are broken up into a chronological set of events. Each video contains several *premise* events, which are explained by one *explanation* event also in the video. AI is tasked with (1) describing these individual *premise* events and (2) formulating a hypothesis in line with the *explanation* event. This same task has also been given to human volunteers. Liang et al. (2022) used several established automated metrics to evaluate AI-generated labels.[14] Overall, AI performed well below the human threshold both in describing *premise* events and formulating hypotheses for *explanation* events. For example, CIDEr evaluation gave the human performance a score of 155.72 for describing premise events and 147.79 for explanatory hypothesis, while the best scores AI received were 38.27 and 30.75, respectively. Interestingly, Liang et al. (2022) did not test only transformer-based NNs but also RNN and transformer+RNN hybrid systems.[15] They found that pure RNN performed worse than pure transformer or hybrid systems. This may suggest that transformer architecture may contain elements which make it better at abductive reasoning than previous architectures. However, further research is required to prove this claim.

This data is philosophically interesting, given the *selective/creative* abduction dichotomy developed in the second section. The reason is that the different datasets and tasks in the literature can be understood as testing these different kinds of abductions, respectively. In NLP, the αNLI dataset tests the linguistic capacity of transformers for *selective* abduction, while αNLG tests creative abductive ability. Meanwhile, in computer vision, we see *localisation* with bounding boxes placed and *comparison* tasks as analogical to visual *selective* abduction. *Localisation* without preplaced boxes and VAR instead test *creative* visual abductive reasoning. Therefore, we can interpret the results analysed as suggesting that transformers can be competent at *selective* abduction and less capable of *creative* abduction. As we saw, transformers seem to, at the very least, closely match human performance at the αNLI dataset and *localisation* task with boxes replaced. In the *comparison* task, transform-

[13] Hessel et al. (2022) remark that a minority of images were not solvable within their testing methodology, so this result applies to solvable instances.

[14] These being BLEU@4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavaie, 2005), ROUGE-L (Lin & Hovy, 2002) and BERTScore (Zhang et al., 2019).

[15] Important differences between RNNs and transformers lie in how they deal with sequence-to-sequence NLP tasks. RNNs depend on *recurrence* in NLP tasks, meaning they relate immediate inputs in one sequence. Once they process one element, they move on to the next, relating it to the hidden state, which they change when processing the next element. Meanwhile, Transformers process input in parallel and use *attention* instead. *Attention* mechanisms allow for relating different input parts across different sequences. For more detailed comparisons, see Merkx and Frank (2021), Karita et al. (2019), and Tang et al. (2018).

ers are further behind, but their performance is still significantly better than chance. Meanwhile, in αNLG, *localisation* without preplaced boxes and in VAR, transformers do much worse. Further, they are better in the linguistic domain in this regard, AND transformers are behind human performance in visual reasoning.[16]

This is important because an optimist about AI capabilities might say that if doctors use primarily *selective* abduction during diagnosis, as Magnani (2001) argued, transformers should be just as capable diagnosticians and hopefully just as able to justify their outputs by providing specific hypotheses for them. It seems that when transformers are provided with a set of hypotheses, they can choose the more plausible hypothesis rather well. However, we would caution against such a conclusion because the deficiencies we discussed are more critical than an optimist would suppose. It is crucial to understand that the capacity for *selective* abduction is too low of a bar for the trustworthy AI that is necessary for healthcare in general. This is due to the arguments that we have made in the second section. *Creative* abduction proves essential when there is no ready-made hypothesis at hand, which means *creative* abduction is useful in situations with a significant amount of uncertainty. That means that if medical AI is deployed in situations of high medical uncertainty, such as the COVID-19 pandemic, transformer-based AI might not be as good at diagnostics. For xAI, this deficiency is likely to prove crucial. As we argued in the first section, situations with high uncertainty are when explanations from AI are more required: we need AI to provide explanations when it arrives at novel conclusions that doctors did not expect. An AI that is not very proficient in *creative* abduction, as transformers are suggested to be by these provisional experiments, could prove relatively inefficient at xAI tasks since it could not generate a new and novel explanation. Instead, it might refer back to already established hypotheses, thus reinforcing the *confirmation* bias (Ghassemi et al., 2021) that arises from AI usage in medicine. This effect would only be compounded by the fact that explanations generated by transformers are not *process-consistent*. Nevertheless, some promising signs are showing with the development of the technology, and we expect this feature to be only a contingency that will be solved in the near future by improving the technology in question.

## 5 Conclusion

In this paper, we have argued that we should reconsider how we grasp explanations in the field of medical xAI. The first section of this paper showed that the current xAI techniques do not alleviate the uncertainty that doctors face in their interactions with AI, creating a TG at two levels that need to be solved. To deal with these queries, we appealed to the *pragmatic turn* in xAI to argue that we need to understand explanations as involving *distinct social and abductive processes that aim to provide answers to contrastive questions.* In order to test how relevant this concept of explanation is

[16] In theory, AI might also have a potential advantage in this area. Human reasoners can get tired or distracted when solving a task. Thus, transformers could be more consistent regarding abductive judgements in the future. However, as was found with ToM tasks, AI performance may lack robustness. Using this advantage is likely to be a matter of further fine-tuning and technical improvements.

in medical xAI, we examined the role that abductive reasoning plays in medicine. We showed that it can be argued to reduce uncertainty in medical decision-making, especially in cases where no ready-made hypotheses are on hand, and that abductive reasoning provides an important first step that further shapes inductive and deductive reasoning processes that may allow the increasing of trust in the doctor-patient relationship that current AI models are missing. While we believe that abduction is no more important than other forms of reasoning, we have concentrated on its explanatory role in this study. Given how under-explored the role of abduction is within the philosophy of medicine, a study of abductive reasoning in medical xAI presents an opportunity to find new ways of solving the problem that the TG presents. As Miller (2019) stressed, abductive reasoning likely plays a vital role in producing explanations. Therefore, it presents itself as an important starting point that allows us to bridge the TG that may arise in traditional medicine. Based on this, we reasoned that if abductive reasoning plays this role, we have to judge the capability of AI on this merit. If AI is found capable in this area, then this capacity could be exploited to improve the xAI capabilities of these systems in healthcare contexts in ways that have not been studied much before.

In the last section of the paper, we engaged with the current state-of-the-art AI systems. We reviewed studies examining transformers' capacity for social and abductive reasoning. We found that explanations generated by LLMs are often *process-inconsistent* and that whatever capacity for social reasoning transformers demonstrate is not robust. More importantly, when we reviewed studies testing the abductive reasoning of transformers, we found that transformers are better at *selective* rather than *creative* abduction, which is significant for their use and implementation in xAI. Suppose AI systems are to provide explanations in situations of great uncertainty, which often arises from novel conclusions generated by AI systems. In that case, they must demonstrate sufficient performance in *creative* abductive reasoning. While in NLP, transformers seem to perform relatively well, in computer vision, they perform worse. This is important for medicine, where AI is primarily used to process multimodal data. We would further like to stress that during our literature review, we learned that AI models are being tested more on *selective* rather than *creative* abduction. When we look at the Allen Institute of AI (n.d.) public leaderboard, we can find that, for example, there are vastly more submissions for the αNLI (121 submissions) leaderboard than the αNLG (8 submissions) leaderboard. We believe that this paper shows the importance of *creative* abduction, and thus, we think that further research is needed to examine *creative* abductive reasoning in AI. [17] We hope this study will lead other xAI and AI researchers to pursue more studies in this direction and moti-

[17] One could raise the question of how the argument presented here can hold any relevance for deductive and inductive AI models per se. The reason why we think abduction is interesting for the Trust Gap is that it can allow the creation of second-order models that can, then, be applied to the deductive or inductive models and help to translate the information that is opaque from a human point of view. At this stage, we are neutral about this translating capacity. However, as we tried to show in the more recent literature on transformers and neural networks, there are some promising signs that, in the near future, we will be able to have conversational agents that can translate, in abductive language, many of the internal processing of current black-model systems. This reinforces one of the claims of this paper: We don't need only Explainable AI; we need Explanatory AI.

vate the exploration of capabilities and the importance of abductive reasoning in this field.[18]

**Acknowledgements** The Authors would like to thank the anonymous reviewers and the Editor in Chief for the useful comments that improved this final version of the paper. SG was a Visiting Researcher at the Robotics Lab (A.S. Centre) at the University of Palermo during the revisions, and presented preliminary versions at the Palermo International Workshop 2024 and the World Congress of Philosophy 2024. JM was an Eramsus+ Visiting Researcher at the Faculty of Letters of the University of Porto during the production of the paper. He presented preliminary versions at the 7th Rebuilding Trust in AI Medicine Online Seminar at the MLAG Seminar at the University of Porto and the World Congress of Philosophy 2024. We would both like to thank the different audiences for their feedback, which inspired some changes in the previous versions of this paper.

**Author Contributions** Authors did an equal amount of work.

**Funding** SG is funded by CEEC Individual Project by FCT 2022.02527.CEECIND, at the Mind, Language and Action Group, Institute of Philosophy, University of Porto (Faculdade de Letras, Via Panorâmica s/n, P-4150-564 Porto, Portugal). JM is funded by Specific Research project "Current and future issues of AI: Need for explainable and ethical AI" supported by the Philosophical Faculty of the University of Hradec Králové in 2024.
Open access funding provided by FCT|FCCN (b-on).

**Data Availability** Not applicable.

## Declarations

**Ethical Approval** Not applicable.

**Informed Consent** Not applicable.

**Statement Regarding Research Involving Human Participants and/or Animals** Not applicable.

**Consent to Participate** Yes.

**Consent to Publish** Yes.

**Competing Interests** Not applicable.



[18] We would like to thank the reviewers for valuable comments that enhanced the overall quality of the arguments presented in this paper.

## References


Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, *28*(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *Ieee Access : Practical Innovations, Open Solutions*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ali, M. J., Hanif, M., Haider, M. A., Ahmed, M. U., Sundas, F., Hirani, A., Khan, I. A., Anis, K., & Karim, A. H. (2020). Treatment options for COVID-19: A review. *Frontiers in Medicine*, *7*, 480. https://doi.org/10.3389/fmed.2020.00480

Aliseda, A. (2006). *Abductive reasoning: Logical investigations into discovery and explanations*. Springer.

Allen Institute of AI (n.d.). *Leaderboards*. (Accessed December 28 2023). https://leaderboard.allenai.org/

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

Andreas, J. (2022). Language models as agent models. In Goldberg, Y., Kozareva, Z., Zhang, Y. (Eds.) *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 5769–5779). https://doi.org/10.18653/v1/2022.findings-emnlp.423

Banerjee, S., Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J, Lavie, A, Lin, C. Voss,C. (Eds.), Proceedings of the ACL on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (pp. 65–72). Association for Computational Linguistics.

Campos, D. G. (2009). On the distinction between Peirce's abduction and Lipton's inference to thebest explanation. *Synthese*, *180*(3), 419–442. https://doi.org/10.1007/s11229-009-9709-3

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., & Choi, Y. S. W. (2020). Y. Abductive Commonsense Reasoning. In International Conference on Learning Representations 2020. https://iclr.cc/virtual_2020/poster_Byg1v1HKDB.html

Blanco, S. (2022). Trust and Explainable AI: Promises and Limitations. *Ethicomp Conference Proceedings*, pp. 245–256.

Briganti, G., & Le Moine, O. (2020). Artificial Intelligence in Medicine: Today and tomorrow. *Frontiers in Medicine*, *7*(27), 509744. https://doi.org/10.3389/fmed.2020.00027

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv*. https://doi.org/10.48550/arXiv.2303.12712

Buckner, C. J. (2024). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, *60*(1), 103111. https://doi.org/10.1016/j.ipm.2022.103111

Campaner, R., & Sterpetti, F. (2023). Abduction, Clinical Reasoning, and Therapeutic Strategies. In Magnani, L. (Ed.), Handbook of Abductive Cognition (pp. 443–465). Springer. https://doi.org/10.1007/978-3-031-10135-9_12

Campos, D. G. (2009). On the distinction between Peirce's abduction and Lipton's inference to the.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical.

Chaves, A. P., & Gerosa, M. A. (2020). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, *37*(8), 729–758. https://doi.org/10.1080/10447318.2020.1841438

Chiffi, D. (2021). *Clinical reasoning: Knowledge, uncertainty, and values in Health Care*. Springer.

Chiffi, D., & Andreoletti, M. (2023). Introduction to Abduction and Medicine: Diagnosis, Treatment, and Prevention. In L. Magnani (Ed.), Handbook of Abductive Cognition (pp. 443–465). Springer. https://doi.org/10.1007/978-3-031-10135-9_83

Consolandi, M., Martini, C., Reni, M., Arcidiacono, P. G., Falconi, M., Graffigna, G., & Capurso, G. (2020). COMMUNI. CARE (COMMUNIcation and patient engagement at diagnosis of pancreatic CAncer): Study protocol. *Frontiers in Medicine*, *7*, 134. https://doi.org/10.3389/fmed.2020.00134

Dai, Y., Gao, Y., & Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, *11*(8), 1384. https://doi.org/10.3390/diagnostics11081384

De Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, *10*, 3061. https://doi.org/10.3389/fpsyg.2019.03061

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Bursten, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Human Language Technologies, Vol. 1, pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*. https://doi.org/10.48550/arXiv.2010.11929

Douven, I. (2021). Abduction. In Zalta, E. N. (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2021 ed.). Stanford University. https://plato.stanford.edu/archives/sum2021/entries/abduction/

Durán, J. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*, 103498. https://doi.org/10.1016/j.artint.2021.103498

Elton, D. (2020). *Self-explaining AI as an alternative to interpretable AI* arXiv. https://doi.org/10.48550/arXiv.2002.05149

Eriksson, K., & Lindström, U. (1997). Abduction—a way to deeper understanding of the world of caring. *Scandinavian Journal of Caring Sciences*, *11*(4). https://doi.org/10.1111/j.1471-6712.1997.tb00455.x. 195 – 198.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Goddard, K., Roudsari, A., & Wyatt, J. (2021). Automation bias a hidden issue for clinical decision support system use. *Studies in Health Technology and Informatics*, *164*, 17–22. https://doi.org/10.3233/978-1-60750-709-3-17

Grangea, J., Princisb, H., Kozlowskib, T., Amadou-dioffo, A., Wu, J., Hicks, Y., & Johansen, M. (2022). XAI & I: Self-explanatory AI facilitating mutual understanding between AI and human experts. *Procedia Computer Science*, *207*, 3600–3607. https://doi.org/10.1016/j.procs.2022.09.419

Gungov, A. L. (2018). The ampliative leap in diagnostics: The advantages of abductive inference in clinical reasoning. *History of Medicine*, *5*(4), 233–242. https://historymedjournal.com/wp-content/uploads/volume5/number4/1.Gungov.pdf

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., & Shen, D. (2023). Transformers in medical image analysis. *Intelligent Medicine*, *3*(1), 59–78. https://doi.org/10.1016/j.imed.2022.07.002

Hessel, J., Hwang, J. D., Park, J. S., Zellers, R., Bhagavatula, C., Rohrbach, A., Saenko, K., & Choi, Y. (2022). The abduction of sherlock holmes: A dataset for visual abductive reasoning. In Avidan S., Brostow G., Cissé M., Farinella G. M. (Eds.), Computer Vision – ECCV 2022 (vol. 13696, pp. 558–575). Springer. https://doi.org/10.1007/978-3-031-20059-5_32

Hoffman, R. R., Clancey, W. J., & Mueller, S. T. (2020). Explaining AI as an exploratory process: The Peircean abduction model. arXiv. https://doi.org/10.48550/arXiv.2009.14795

Holm, S. (2023). On the justified use of AI decision support in evidence-based medicine: Validity, Explainability, and responsibility. *Cambridge Quarterly of Healthcare Ethics*, 1–7. https://doi.org/10.1017/S0963180123000294

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15565–15575). IEEE Computer Society. https://doi.org/10.1109/CVPR52688.2022.01512

Jentzsch, S. F., Höhn, S., & Hochgeschwender, N. (2019). Conversational interfaces for explainable AI: a human-centred approach. In Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (Eds.), Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019 (pp. 77–92). Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_5

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., & Zhang, W. (2019). A comparative study on transformer vs rnn in speech applications. *2019 IEEE Automatic Speech Recognition and understanding workshop (ASRU)* (pp. 449–456). IEEE. https://doi.org/10.1109/ASRU46091.2019.9003750

Karlsen, B., Hillestad, T. M., & Dysvik, E. (2020). Abductive reasoning in nursing: Challenges and possibilities. *Nurse Inquiry*, *28*. https://doi.org/10.1111/nin.12374

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pp. 169–175.

Khan, S., Naseer, M., Hayat, M., W Zamir, S., S Khan, F., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

Kim, S., Huh, I., Park, Y., & Lee, S. (2022). Designing a pragmatic explanation for the XAI system based on the user's context and background knowledge. In Nam, C. S., Jung J., Lee, S. (Eds.), Human-Centered Artificial Intelligence (pp. 117–125). Academic Press. https://doi.org/10.1016/B978-0-323-85648-5.00012-8

Kosinski, M. (2023). Theory of mind might have spontaneously emerged in large language models. arXiv.https://doi.org/10.48550/arXiv.2302.02083

Lebovitz, S. (2020). Diagnostic doubt and artificial intelligence: An inductive field study of radiology work. In *Proceedings of the 40th International Conference on Information Systems* (Vol. 7, pp. 5385–5401). Curran Associates, Inc.

Liang, C., Wang, W., Zhou, T., & Yang, Y. (2022). Visual Abductive Reasoning. In Proceedings of the.

Lim, B. Y., Yang, Q., Abdul, A. M., & Wang, D. (2019). Why these explanations? Selecting intelligibility types for explanation goals. In Trattner C. Parra, D., Riche N. (Eds.), Joint Proceedings of the ACM IUI 2019 Workshops (vol. 2327). https://ceur-ws.org/Vol-2327/

Lin, C. Y., & Hovy, E. (2002). Manual and automatic evaluation of summaries. In Proceedings of the ACL-02 Workshop on Automatic Summarization (pp. 45–51). Association for Computational Linguistics. https://doi.org/10.3115/1118162.1118168

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A survey of transformers. *AI Open*, *3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, *27*, 247–266. https://doi.org/10.1017/S1358246100005130

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2012). Explanation and Abductive Inference. In Holyoak K. J., Morrison R. G. (Eds.), The Oxford Handbook of Thinking and Reasoning, Oxford Library of Psychology, 260–276. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*(6), 975–995. https://doi.org/10.1007/s11229-011-0054-y

Magnani, L. (2001). *Abduction, reason and science: Processes of Discovery and Explanation*. Springer.

Martini, C. (2023). Abductive Reasoning in Clinical Diagnostics. In Magnani, L. (Ed.), *Handbook of Abductive Cognition* (pp. 467–479). Springer. https://doi.org/10.1007/978-3-031-10135-9_13

Medianovskyi, K., & Pietarinen, A. (2022). On explainable AI and Abductive Inference. *Philosophies*, *7*(2), 35. https://doi.org/10.3390/philosophies7020035

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In Chersoni, E., Hollenstein, N., Jacobs, C., Oseki, Y., Prévot, L., Santus, E. (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2021*, pp. 12–22). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, *36*, e14. https://doi.org/10.1017/S0269888921000102

Miller, T. (2023). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support. In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 333–342). https://doi.org/10.1145/3593013.3594001

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv*. https://doi.org/10.48550/arXiv.1712.00547

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 279–288). https://doi.org/10.1145/3287560.3287574

Nguyen, V. B., Schlötterer, J., & Seifert, C. (2023). From Black Boxes to Conversations: Incorporating XAI in a Conversational Agent. In Longo, L. (Ed.), Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science (vol. 1903, pp. 71–96). Springer. https://doi.org/10.1007/978-3-031-44070-0_4

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education*, *39*(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, *24*(1), 13. https://doi.org/10.1007/s10676-022-09632-3

OpenAI (2023). *GPT-4 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Páez, A. (2019). The pragmatic turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, *29*, 441–459. https://doi.org/10.1007/s11023-019-09502-w

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E.,Lin, D. (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Peirce, C. S. (1931–1958). Collected papers of Charles Sanders Peirce, Vols. 1–6, Hartshorne, C., Weiss, P. (Ed.); Vols. 7–8, Burks, A. W. (Ed.), Harvard University Press.

Pesapane, F., Codari, M., & Sardanelli, F. (2018). Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, *2*(1), 1–10. https://doi.org/10.1186/s41747-018-0061-6

Picard, R. W. (2000). *Affective Computing*. MIT Press.

Piccialli, F., Somma, V. D., Giampaolo, F., Cuomo, S., & Fortino, G. (2021). A survey on deep learning in medicine: Why, how and when? Information Fusion, 66, 111–137. https://doi.org/10.1016/j.inffus.2020.09.006

Pietarinen, A. V., & Bellucci, F. (2014). New light on Peirce's conceptions of retroduction, deduction, and scientific reasoning International Studies in the Philosophy of Science, 28 (4), 353–373.https://doi.org/10.1080/02698595.2014.979667

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 4218–4227). PMLR. Retrieved June 10 2024, from https://proceedings.mlr.press/v80/rabinowitz18a.html

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* OpenAI. Retrieved December 28 2023, from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ramoni, M., Stefanelli, M., Magnani, L., & Barosi, G. (1992). An epistemological framework for medical knowledge-based systems. *IEEE Transactions on Systems Man and Cybernetics*, *22*(6), 1361–1375. https://doi.org/10.1109/21.199462

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngomo, A. N., Schulte, C., Wagner, W. H., & Wrede, P., B (2021). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Sap, M., Le Bras, R., Fried, D., & Choi, Y. (2022). Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In Goldberg, Y., Kozareva, Z. Zhang, Y. (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 3762–3780). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.248

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 1–59. https://doi.org/10.1007/s10618-022-00867-8

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). *Clever hans or neural theory of mind? Stress testing social reasoning in large language models*. arXiv. https://doi.org/10.48550/arXiv.2305.14763

Sovrano, F., & Vitali, F. (2022). Explanatory artificial intelligence (YAI): Human-centered explanations of explainable AI and complex data. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-022-00872-x

Stanley, D. E., & Nyrup, R. (2020). Strategies in abduction: Generating and selecting diagnostic hypotheses. *The Journal of Medicine and Philosophy*, *45*(2), 159–178. https://doi.org/10.1093/jmp/jhz041

Tang, G., Müller, M., Gonzales, A. R., & Sennrich, R. (2018). Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., (Eds.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4263–4272). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1458

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. arXiv.https://doi.org/10.48550/arXiv.2302.08399

van Duijn, M. J., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M. R., & van der Putten, P. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7–10 on Advanced Tests. In Jiang, J., Reitter, D., Deng, S. (Eds.), Proceedings of the 27th Conference on Computational Natural Language Learning (pp. 389–402). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.25

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000–6010). Curran Associates Inc.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on Computer vision and Pattern Recognition (pp. 4566–4575). https://doi.org/10.1109/CVPR.2015.7299087

Veen, M. (2021). Creative leaps in theory: The might of abduction. *Advances in Health Sciences Education*, *26*, 1173–1183. https://doi.org/10.1007/s10459-021-10057-8

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839. https://doi.org/10.1016/j.artint.2022.103839

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 12113–12132. https://doi.org/10.1109/TPAMI.2023.3275156

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. arXiv. https://doi.org/10.48550/arXiv.1904.09675

Zhang, Z., Wang, S., Xu, Y., Fang, Y., Yu, W., Liu, Y., Zhao, H., & Zeng, Z. C. (2022). M. Task Compass: Scaling Multi-task Pre-training with Task Prefix. In Goldberg, Y., Kozareva, Z. Zhang, Y. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 5671–5685). https://doi.org/10.18653/v1/2022.findings-emnlp.416

Zhang, H., Ee, Y. K., & Fernando, B. (2024). A region-prompted adapter tuning for visual abductive reasoning. arXiv. https://doi.org/10.48550/arXiv.2303.10428

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., & Wen, J. (2023). *A Survey of Large Language Models* arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zheng, X. (2023). Joint Abductive Generation and Discrimination via Cycling Reasoner. Github. Retrieved December 28, 2023, from https://github.com/MrZhengXin/abductive_reasoning_cycle/blob/main/Joint_Abductive_Generation_and_Discrimination_via_Cycle_Reasoner.pdf

## Authors and Affiliations

**Steven S. Gouveia[1,2] · Jaroslav Malík[3]**

✉ Steven S. Gouveia
stevensequeira92@hotmail.com

[1] Mind, Language and Action Group of the Institute of Philosophy of the University of Porto, Porto, Portugal

[2] Honorary Professor Faculty of Medicine, Andrés Bello University, Santiago, Chile

[3] Philosophy Department, University of Hradec Králové, Hradec Králové, Czech Republic