Explanatory Dialogues and Digital Medicine

Bodlović, Petar; Lewiński, Marcin; Cabrio, Elena; Villata, Serena; Boogaart, Ronny; Garssen, Bart; ... ; Reuneker, Alex

Boogaart, R., Garssen, B. Jansen, H., Leeuwen, M. van, Pilgram, R. & Reuneker, A. (2024).
Proceedings of the Tenth Conference of the International Society for the Study of Argumentation.
Sic Sat: Amsterdam.

Explanatory Dialogues and Digital Medicine

PETAR BODLOVIĆ, MARCIN LEWIŃSKI, ELENA CABRIO & SERENA VILLATA

NOVA Institute of Philosophy
NOVA University Lisbon
Portugal
pbodlovic@fcsh.unl.pt

NOVA Institute of Philosophy
NOVA University Lisbon
Portugal
m.lewinski@fcsh.unl.pt

Université Côte d'Azur
Inria, CNRS, I3S
France
elena.cabrio@univ-cotedazur.fr

Université Côte d'Azur
Inria, CNRS, I3S
France
villata@i3s.unice.fr

ABSTRACT: To what extent can the model of explanatory dialogue elucidate the interaction between the Doctor and diagnostic AI, i.e., Explainable Artificial intelligence (XAI)? We argue that explanatory dialogue models do not make an optimal framework for analyzing and evaluating the Doctor-XAI interaction. This is because this interaction, typically, does not aim at transferring *understanding*, a main goal of explanatory dialogue.

KEY WORDS: artificial intelligence, clarification, defeaters, Douglas Walton, epistemic luck, explanatory dialogues, explainable artificial intelligence, justification, medicine, understanding

1. INTRODUCTION

Explanations are ubiquitous in the medical domain. While doctors try to identify a disease that explains the patient's symptoms (Josephson & Josephson, 1994), researchers often focus on etiology and identify causes that explain diseases (Marcum, 2008; Dragulinescu, 2016). Accordingly, scholars recognize that the *inference to the best explanation* (henceforth, IBE) is essential to medical practice. In a broad sense, IBE amounts to generating, justifying, and accepting a hypothesis *H* because *H* provides the best explanation of data (Harman, 1965; Douven, 2021). For instance, a doctor may conclude "The patient has hemolytic anemia" because this type of illness— together with background information (age, gender, medical history, etc.)—best explains the patient's symptoms (e.g., dark urine).

Explanations are frequently used in dialogues: in response to the User's question ("Why *p*?"), the Explainer explains *p* for the sake of clarifying (Cawsey, 1992) or transferring her understanding (Walton, 2011). In fact, since scholars often define

explanation in terms of its communicative goal,[1] and the User's interests, commitments, and criticisms determine if the explanation achieves this goal, explanations might be portrayed as inherently dialogical. In that case, dialectics provides essential tools for studying explanations. Applying a normative model of explanatory dialogue should contribute to the analysis and evaluation of a particular explanatory interaction.

Explanatory dialogues in medicine include different parties: a doctor might be explaining something to a patient, family member, another doctor, or she might demand an explanation from some Artificial Intelligence System (henceforth, AI). The usage of AI in diagnostics and medical prediction has increased recently but, at the same time, software has become extremely complex and started to lack interpretability. Namely, when the AI system generates an output—e.g., a correct diagnosis that "The patient has hemolytic anemia"—doctors and software engineers find it difficult to understand how, or on what grounds, the system produces such output (Holzinger et al., 2019). Since treating them as 'black boxes' is epistemically and morally problematic, AI systems must become explainable. Making AI transparent and interpretable by clarifying a software's training, database, restrictions, correlations, inferences, etc. is the main goal of *explainable AI* (henceforth, XAI), an emerging field in AI.

This paper's main question is: Does the model of explanatory dialogue—proposed, for instance, by Douglas Walton (2011)—elucidate the interaction between the doctor and XAI? After all, the Doctor appears to assume the User's role. She faces something she does not understand and demands an explanation from XAI (the Explainer): "Why output (hemolytic anemia, rather than…), based on the input (patient's symptoms, biographical information, etc.)?" Surprisingly, we argue that explanatory dialogue models, although helpful to some degree, do not make an optimal framework for analyzing and evaluating the Doctor-XAI interaction. This is because such interaction often does not aim to transfer *understanding*—the main goal of explanatory dialogue—but rather justification, knowledge, or truth. Consequently, studying the Doctor-XAI interaction from an argumentative viewpoint, e.g., that of a critical discussion (van Eemeren & Grootendorst, 2004) or inquiry (Walton & Krabbe, 1995), seems more promising.

The paper proceeds as follows. Section 2 provides an overview of explanation studies in argumentation theory. Section 3 describes the features of explanatory dialogues. In Section 4, we introduce an example from medical diagnostics and explain what motivates the development of explainable artificial intelligence (XAI). Section 5 shows that although the Doctor-XAI interaction might be initiated by a clarification request, it assumes more fundamental goals. In Section 6, we sketch several arguments why the Doctor-XAI interaction does not aim at understanding and, consequently, should not be primarily analyzed in terms of an explanatory dialogue. Section 7 summarizes the main results.

2. EXPLANATIONS IN ARGUMENTATION THEORY

Explanations raise many kinds of questions. First, theorists focus on *conceptual demarcations* and questions like: What distinguishes an 'explanation' from 'reasoning' and 'argument' (Wright, 2002; McKeon, 2013; Dufour, 2017; Govier, 2018)? On

[1] For instance, an explanation is meant to "make clear to the hearer some complex piece of knowledge" (Cawsey 1992: 5), "achieve the state where hearer knows" (Moore 1995: 10), or "transfer … understanding from the one party to the other" (Walton 2011: 356).

standard accounts, both argument and explanation instantiate reasoning, but while arguments "provide evidence for accepting a doubted conclusion," explanations "provide a cause for an already accepted conclusion" (Mayes, 2010, p. 95). So, for instance, the Doctor demands an explanation by asking "Why does the patient have dark urine?" and, presumably, an argument by asking "What makes a hemolytic anemia the best diagnosis?"

Second, scholars study the explanation's formal dimensions. Their research centers on developing *argument schemes* (with the list of critical questions) associated with paradigmatic explanatory inferences, e.g., IBE (Walton et al., 2008; Wagemans, 2016; Yu & Zenker, 2018; Olmos 2021). Third, *pragmatic research* focuses on defining the speech act of explaining (Gaszczyk, 2023) and studying its applications in communication. For instance, transferring understanding—as the essential function of explaining (Lipton, 2004; Grimm, 2010)—becomes critical when the Challenger cannot figure out the Proponent's commitments and, consequently, fruitfully participate in a discussion. In such circumstances, explaining facilitates communication as a 'local' move made in a broader argumentative context (van Laar and Krabbe, 2013).

Studying explanations as moves in argumentative dialogues, however, is different from studying *explanatory dialogues*, i.e., dialectical procedures that systematically promote the transfer (or enhancement) of understanding.

3. EXPLANATORY DIALOGUES

Cawsey (1992), Moore (1995), and Walton (2004, 2005, 2011) have developed models of explanatory dialogues, thereby explicitly engaging in the "explanatory dialectic" (Rescher, 2007, p. 48). Cawsey and Moore emphasize that, sometimes, the User does not understand the computer's explanation, so the computer must respond to her feedback. We summarize Walton's model below.

First, at the beginning of the explanatory dialogue, the User lacks understanding: she encounters "an anomaly in the account" (Walton, 2011, p. 350), surprising data that does not cohere with her expectations based on 'how things normally work' (e.g., dark urine). Second, the User assumes that the Explainer understands the anomaly, and requests an explanation ("Why is the urine dark?"). The goal of explanatory dialogue is "a transfer of understanding from the one party to the other" (2011, p. 356) and, for this to happen, the User "is supposed to understand all the premises in the chain of reasoning" (2005, p. 81). Third, the Explainer offers an explanation and the User either accepts it or requests further reasons or clarifications. At this point, an explanatory dialogue may shift into an argumentative one, such as examination (2011). Finally, the understanding is transferred if the Explainer responds to all questions and challenges, and the User does not have further demands. We can present this simplified version of Walton's model in a systematic dialectical fashion.

INITIAL SITUATION	(1) The User is puzzled by some anomaly she does not (but wants to) understand, (2) The Explainer has understanding.
COLLECTIVE GOAL	Transferring understanding from the Explainer to the User.
OPENING STAGE	Agreeing on roles, and starting points (propositions both parties *understand*).

EXPLANATION STAGE	The Explainer responds to the User's explanation request [*possible shifts to other dialogue types*].
CLOSING STAGE	The explanation is judged to be successful or not (i.e., to transfer understanding or not).

Figure 1. Walton's (2004, 2005, 2011) Model of Explanatory Dialogue

Does this model elucidate the Doctor-XAI interaction? At the beginning of a diagnostic procedure, the Doctor gathers data: the patient's biographical information, medical history, results of physical examination, laboratory tests, etc. Consider the real-life example from *casiMedicos*:[2]

> A 32-year-old woman with cerebral palsy from childbirth came to the emergency department for a few days of dark urine associated with an episode of high fever and dry cough. On admission, the CBC showed 16900 leukocytes/mm3 (85% S, 11% L, 4% M) ... In the biochemistry LDH 2408; bilirubin 6.8 mg/dl, … The morphological study of blood showed macrocytic anisocytosis with frequent spherocytic forms and polychromatophilia without blasts. The irregular antibody study is positive in the form of panagglutinin…

Next, the Doctor feeds the AI system with this data (inputs), gets the diagnosis "The patient has hemolytic anemia associated with respiratory infection" (output), and requests an account of why these inputs generate this output. So, apparently, the Doctor assumes the User's role and eventually interacts with the XAI (assuming the Explainer's role).

4. EXPLAINABLE ARTIFICIAL INTELLIGENCE

As Nyrup & Robinson (2022, p. 12) put it, "medical AI systems are developed to take on increasingly critical roles in assisting with clinical reasoning tasks such as diagnosis, prognosis, and treatment decisions." The problem is that their "applications became increasingly successful but increasingly opaque" (Holzinger et al., 2019, p. 1): it became difficult to spell out the grounds and procedures the AI uses to transform inputs into outputs. In effect, XAI is developed to enhance the "transparency and traceability of statistical black-box machine learning methods" (Theunissen & Browning, 2022, p. 2) and tackle "multiple layers of opacity" (Holzinger et al., 2019, p. 1). But what does the 'opacity' mean? What conditions reduce explainability and why is it important to overcome them?

Nyrup & Robinson (2022) argue that explainability is not well-defined. They understand it as AI's relational property—depending on context, audience, and purpose—and identify features that reduce it. First, explainability is reduced by *secrecy*. For instance, some software cannot convey information about their inner states: they are simply not designed to formulate an explanation, explicitly, on demand. Second, the audience can lack the vocabulary to interpret an explanation, thereby reducing AI's explainability (*technical literacy*): often, laymen do not understand technical concepts and formalisms. Third, the system's *complexity* can render understanding cognitively demanding, despite the User's technical literacy.

[2] *casiMedicos* is a project where collaborators attempt to make various medical data more accessible to public (see https://www.casimedicos.com/). For instance, in one part of the project, voluntary medical doctors explain examples from MIR exams. We took our example from this data set.

> [A]dvanced machine learning models, … can take hundreds or thousands of input variables and use these to calculate a highly nonlinear and non-monotonic function (Selbst & Barocas, 2018, 1094-96), meaning that there are no simple, overall rules for whether increasing a given input variable will increase or decrease (and by how much) the probability of a given decision. (Nyrup & Robinson, 2022, p. 8)

Fourth, sometimes, the User cannot translate AI's computational states into human concepts (*semantic mapping*). Suppose that the medical AI 're-groups' original inputs into larger units that seem entirely random to the human mind: e.g., {dark urine, mother of two, macrocytic anisocytosis}, {high fever, married, bilirubin 6.8 mg/dl}. The User might observe how such units interact and understand their syntactic behavior, but cannot comprehend their semantic meanings. Finally, the AI system can draw seemingly irrelevant inferences. Suppose AI somehow (correctly) infers "The patient has hemolytic anemia" from "The patient had a traffic accident twenty years ago." Since the User cannot integrate this inference into her *domain knowledge*—i.e., how things normally (causally) work—the AI lacks explainability. So, according to Nyrup & Robinson's *Explanatory Pragmatism*, both the AI's and the User's features contribute to defining and enhancing explainability.

Explainability is important for epistemic, pragmatic, and ethical reasons. Suppose the Doctor makes a diagnosis *D* based on evidence *E*. If AI suggests a conflicting diagnosis *D**, the Doctor cannot assess its epistemic significance unless AI's reasoning is explainable: the Doctor must know whether AI has taken *E* into account to assess whether *D** defats *D*, or, perhaps, *E* defeats *D**. Moreover, to decide whether to apply AI in medical practice, its training must be transparent: if some diagnostic software was trained on the data concerning white males, its applications are risky when patients are, predominantly, black females (Nyrup & Robinson, 2022, pp. 10-12). Finally, XAI is motivated by ethical concerns. In high-risk domains (e.g., medicine, law, or finance), the Users are entitled to explanations (Cappelen & Dever, 2021, p. 25). Explainability enables autonomous decision-making—governed by the stakeholder's values—about matters that directly influence their lives (e.g., "To take an additional round of chemotherapy, or not?"), thereby reducing the threat of paternalism (Theunissen & Browning, 2022, pp. 7-8).

Studying the Doctor-XAI interaction within the pragmatic dialectical framework seems perfectly reasonable. Both dialecticians and XAI researchers stress the essential roles of audience and the contextual goals in assessing dialogue moves. And since XAI is all about understanding, interpretability, and explainability, applying the normative model of explanatory dialogue seems like an obvious choice. But is it an optimal choice? Do diagnosticians seek understanding (transferred via explanations), or do they demand clarifications, justifications, or knowledge (transferred via arguments)? Since the evaluation of some contribution depends on the dialogue's 'collective goal,' we must closely examine this question.

5 THE GOAL OF XAI: CLARIFICATION?

Clarifications improve the User's understanding, but, unlike explanations, do not enhance understanding of events, or facts (i.e., anomalies, such as the dark urine) but rather obscure expressions, or speech acts (e.g., unclear concepts, or ambiguous utterances) (Walton, 2007, 2011; van Laar & Krabbe, 2013). While explanations

address puzzling states of the world, clarifications address obscure meanings of the words. Accordingly, the collective goal of clarification dialogue is to help the Hearer to understand an obscure or problematic utterance of the Speaker (Walton, 2007).

AI's ability to clarify utterances improves explainability. Suppose AI supports its diagnosis by referring to its sophisticated statistical methods. However, this explanation contains technical AI jargon—e.g., 'noise-aware machine learning'—that the Doctor is unfamiliar with. To enhance explainability, first, AI may translate 'noise-aware machine learning' into a more conventional expression the Doctor is familiar with (such as 'measurement error') (Faes et al., 2022). Then, AI might clarify the meaning of this expression: "This diagnostic software is sensitive to the possibility that measured values of original variables do not correspond to their true values due to observational, environmental, or instrumental factors generating measurement errors." Accordingly, AI's clarificatory abilities might reduce secrecy and technical illiteracy.

However, clarifications are not enough. Even when all utterances are clear, AI's reasoning might be too complex, or based on correlations that seem irrelevant to humans (domain knowledge). Crucially, since the Doctor is interested in facts and events—causal or symptomatic relations in the real world relevant for understanding the patient's condition—rather than linguistic meanings, clarifying cannot be the ultimate goal of the Doctor-XAI interaction. Even if necessary, clarifying is not sufficient to assess some diagnosis or understand why AI selects it.

If understanding what something means is an auxiliary goal in the Doctor-XAI dialogue, what is the essential one? Does the Doctor interact with XAI to *understand why* "The patient has hemolytic anemia?", or, more precisely, why "AI inferred 'The patient has hemolytic anemia' [instead of some other diagnosis]?" Or does she want to *know why* (or if) the diagnosis is true or was inferred by AI? Is the Doctor-XAI dialogue primarily explanatory or argumentative?

6. THE GOAL OF XAI: TRANSFERRING UNDERSTANDING, JUSTIFICATION, OR TRUTH?

Figure 2 shows how the dialogue between the Doctor and AI naturally progresses.[3] Initially, the Doctor takes information about the patient for granted but does not understand why some of it is true. Symptoms *S* represent an anomaly, and she wants to understand why the patient has them.[4] So, the Doctor requests the explanation at t_1, and AI responds with a diagnosis *D* at t_2.

For now, the dialogue goal is explanatory and corresponds to Walton's model: AI is transferring its understanding of why *S*. But since the Doctor wants the best available explanation, she tests *D*'s acceptability. For instance, she might suspect *D** is an accurate diagnosis and, at t_3, introduce the so-called "argued challenge" (van Laar & Krabbe, 2013, p. 206): "Why autoimmune hemolytic anemia, instead of acute leukemia? Acute anemia also explains dark urine, fever, and dry cough."[5] To be adequately explainable in such circumstances, AI must offer reasons that justify

[3] This is not, by any means, the complete dialogue profile of the Doctor-XAI interaction. The dialogue can also take other, alternative routes.
[4] That symptoms are usually "those findings for the case that show abnormal values" (Josephson & Josephson, 1994, p. 9) fits nicely with the notion of 'anomaly.'
[5] Argued challenges expose the *contrastive nature* of statements, explanations, and arguments (see Dretske, 1972; Lipton, 2004).

selecting *D* over *D**, e.g. (from t_4 to t_6): "Autoimmune hemolytic anemia explains all symptoms. Leukemia neither explains the absence of blasts in the blood, nor the presence of panagglutinin." So, the explanatory dialogue shifts into argumentation: explaining symptoms in terms of a diagnosis turns into justifying diagnosis in terms of symptoms.

Time	**Dialogue party**	**Type of dialogue move**	**Dialogue move**
t_1	The User (Doctor)	Why$_{(exp)}$ *S*? *Request for explanation*	Given biographical information [age, gender], medical history [cerebral palsy], etc., why does the patient have these symptoms [dry cough, high fever, dark urine; blood test results, etc.]?
t_2	The Explainer (AI)	Because [diagnosis] *D*. *Explanation*	Because of autoimmune hemolytic anemia associated with respiratory infection.
t_3	The User (Doctor)	Why$_{(arg)}$ *D*? (What about *D**?) *Request for justification*	Why anemia? What about acute leukemia? (Acute leukemia also explains dark urine, fever, and dry cough.)
t_4	The Explainer **(XAI)**	Because *E*. *Argument*	Because anemia is the best explanation for observed symptoms. *Inference to the Best Explanation*
t_5	The User (Doctor)	Why$_{(arg)}$ *E*? *Request for justification*	Why is anemia the best explanation, i.e., a better explanation than acute leukemia?
t_6	The Explainer **(XAI)**	Because *A*. *Argument*	Because while anemia explains all symptoms, leukemia neither explains the absence of blasts in the blood, nor the presence of panagglutinin. *Inference to the Best Explanation, continued…*

Figure 2. The Doctor-AI (XAI) interaction: natural progression

To answer our main question—"Can the model of explanatory dialogue elucidate the interaction between the doctor and XAI?"—we must focus on the dialogue's second part. Namely, XAI is activated only at t_4—prompted by the Doctor's argued, contrastive challenge at t_3. Before, AI acts as a 'black box:' it does not explain why symptoms *S* imply diagnosis *D* (and not *D**). So, we can formulate our research question, as follows: Can the model of explanatory dialogue elucidate the sequence of Doctor-XAI interaction from t_3 to t_6?

As we have seen, the sequence from t_3 to t_6 is genuinely argumentative. Explanatory considerations come to the fore but perform a justificatory function.[6] Granting that the dialogue is explanatory in this sense, our essential question—in its final, most precise formulation—concerns the relevance of the explanatory dialogue's goal: Does the dialogue from t_3 to t_6 aim at *transferring understanding* from XAI to the Doctor? We argue that, typically, it does not. But before sketching our arguments, we must introduce the last piece of the puzzle: the meaning of 'understanding.'

6.1 What is understanding?

Scholars define understanding as cognitive success (Elgin, 2007) that "essentially involves an element of 'grasping' or 'seeing'" (Grimm, 2010, p. 342). To understand why *D*, one must grasp "*S*, therefore *D*," i.e., have this inference under cognitive control (Hills, 2016). Many cognitive abilities indicate cognitive control. For instance, grasping some general principle entails one's ability "to apply the principle to particular situations" (Grimm, 2010, p 341). However, our diagnostic dialogue (from t_3 to t_6) focuses on grasping a particular diagnosis rather than some general principle (such as natural law): the Doctor wonders why a particular patient should be diagnosed with a particular disease, given her unique biography, history of illness, and a set of symptoms.

Grasping such inferences requires the ability to infer correct conclusions about similar cases, or flawlessly manipulate the initial variables (Hills 2009, 2016). As Grimm puts it, one must be able to "anticipate how changes in the value of one of the variables … would lead to (ceteris paribus) a change in the value of another variable" (2010, pp. 340-341). For instance, to understand "Eating animals is wrong because of the suffering of animals under modern farming methods" an agent must reason correctly about cases where animals do not suffer or are not exposed to farming methods (e.g., wild animals, fish, etc.) (Hills, 2009, p. 100). In this sense, understanding "*S*, therefore *D*" assumes one's "capacity to answer new questions" (Scriven 1972, p. 32).

6.2 Understanding is irrelevant, extra-dialogical, or secondary

Is the purpose of communicating with a diagnostic AI system—and XAI as its internal part—to enable doctors to reason in similar cases, answer new questions, or manipulate variables?

This seems counterintuitive. If the case concerns a 32-year-old woman with cerebral palsy, dark urine, and dry cough, then the Doctor is interested in diagnosing *this* woman, with *this* medical history, and *these* symptoms. She wants to know why AI transforms these specific inputs into *this* specific output (autoimmune hemolytic anemia associated with respiratory infection). Consider the following questions:

[6] AI *justifies* accepting *D* by showing that *D* is the best explanation of *S*.

- Would the diagnosis be different if the patient were also a drug addict a long time ago?
- Would the diagnosis be different if the patient also had night sweats?

The Doctor's ability to answer these questions demonstrates understanding, but it is difficult to see how adequate management of false and random variables contributes to diagnosing the actual patient. If the patient was not a drug addict, then grasping the connection between addiction and anemia—i.e., understanding—is diagnostically useless and cannot be a goal of our Doctor-XAI interaction.

Notice, however, that we created the previous case by adding a random (and false) variable to the actual case. What about manipulating existing (non-random, and true) variables? Consider the next questions:

- Would the present diagnosis be different if the patient were a 32-year-old man, or 62-year-old woman, or did not have cerebral palsy, or did not report high fever, or…?
- Would the existing set of symptoms be different if acute leukemia were the correct diagnosis, or…?

The ability to answer these questions might be crucial to diagnosing an actual patient. Sometimes, inferring a correct diagnosis in similar cases—e.g., when the patient is a 32-year-old *man*—proves the Doctor's ability to evaluate the epistemic relevance (and significance) of available evidence. That is, suppose that, when variables change, the Doctor selects the implausible diagnoses. This indicates her general incompetence in evaluating a causal significance of age, sex/gender, and cerebral palsy—or symptomatic significance of high fever—for having hemolytic anemia. This general incompetence might entail a particular one: the Doctor will not be able to assess evidence about the actual patient and, as a result, select the epistemically justified diagnosis. Furthermore, to select the most plausible diagnosis, the Doctor must eliminate alternative hypotheses, and this also requires reasoning in similar cases (Khalifa, 2017, p. 56). She might assume leukemia is the right diagnosis, identify expected symptoms, and reject the hypothesis by showing that these symptoms do not correspond to actual ones. Hence, grasping is sometimes diagnostically relevant.[7]

However, per se, this does not make understanding the goal of the Doctor-XAI interaction. If the dialogue were about transferring understanding, the Doctor would test the AI's grasping abilities by asking questions, such as: "What if the patient were a 32-year-old man?" Such questions are naturally asked during the software's *training*—because engineers seek to develop reliable software applicable to many patients and circumstances—but seldom arise during our Doctor-XAI interaction. The Doctor is more interested in testing the AI's diagnosis and justifications (by asking critical questions) than examining the AI's grasping abilities. Notice, however, that some questions achieve both: by requesting AI to eliminate an alternative hypothesis, the Doctor tests the diagnosis *and* examines AI's ability to manipulate variables.

Accordingly, our research question does not have a straightforward answer. While some sorts of understanding (1) do not contribute to the Doctor-XAI interaction, other sorts are essential to its success: sometimes, grasping is precisely what the Doctor

[7] This coheres well with our previous conclusion that, in this interaction, explanations provide justifications.

needs to evaluate the available evidence and select the best diagnosis.[8] However, understanding remains secondary: it is legitimate as long as it promotes selecting the most plausible diagnosis.

6.3 Understanding is too epistemically permissive

The understanding's relations with truth, justification, and epistemic luck also challenge the hypothesis that doctors interact with XAI to gain understanding.

First, some scholars believe understanding is non-factive (Elgin, 2007). One might be able to coherently manipulate the variables—while getting supposedly, or approximately accurate results—while using false propositions. For instance, de Regt (2015, p. 3782) argues that Newton's theory of gravitation provides an understanding of gravitational phenomena, although Einstein's theory of general relativity refuted its essential hypothesis: the existence of attractive forces. So, a physicist working within Newton's framework understands both the reality and gravitation theory, although the theory is false and its descriptions are inferior approximations. Similarly, a false diagnosis might produce an understanding of accurate symptoms, and false evidence can produce an understanding of an accurate diagnosis. Acquiring a cognitive state that is, in one important sense, insensitive to truth cannot be the primary goal of the diagnostic dialogue.

Second, even if understanding is factive—i.e., requires grasping connections between true propositions (Grimm 2010)—it is not sensitive to epistemic defeaters.[9] To justify a diagnosis *D*, the AI should recognize and eliminate reasons that (potentially) undermine *D*. However, Hills (2016) argues that understanding is not sensitive to misleading evidence. To reformulate her example, suppose the AI generates accurate diagnosis *D* without taking into account a relevant symptom *M*. If AI added *M* to her set of inputs, the diagnosis would be different (i.e., false). Luckily, *M* is a piece of misleading evidence, and ignoring it led to an accurate diagnosis. In this case, AI has the understanding of why *D*, although producing *D* was epistemically unwarranted: evidence that seemed relevant at the time was set aside. A cognitive state that permits ignoring relevant evidence cannot represent the primary goal of a diagnostic dialogue.

However, the final epistemic 'insensitivity' of understanding—the one concerning epistemic luck—adds an unexpected twist to the story. Surprisingly, it might make understanding a more suitable goal of diagnostic dialogue than knowledge or justification. Let's paraphrase another example from Hills (2016). Suppose AI correctly infers *D* from the particular set of symptoms. However, for any other set of symptoms, AI would select the wrong diagnosis and provide a bad explanation. In this case, AI understands why *D* is true, but lacks justification: in an obvious sense, the true diagnosis was acquired accidentally, by the unsafe process of belief-formation (Pritchard, 2007). But notice that, in a concrete diagnostic case, the Doctor might find the 'lucky' nature of AI's inference completely irrelevant. As we have already seen, the Doctor is primarily interested in getting an accurate diagnosis and strong justification concerning a particular patient (a 32-year-old woman with cerebral palsy, dark urine,

[8] However, that the Doctor's grasping might be essential for the dialogue's success still does not mean that her understanding is dialogically produced. If understanding *does not result* from the interaction, but is triggered by it, or simply applied to it, then it cannot represent the dialogue's final goal (in a traditional sense).

[9] Defeater-sensitivity is crucial for justification, and "justified belief has an intimate link to true belief" (Goldman, 2003, p. 62). Obviously, a diagnosis is more likely to be true if it is strongly justified, and the main goal of diagnostic procedure is to select the true diagnosis.

etc.), and, in this respect, AI does the job perfectly. Its inference is accurate, based on strong medical grounds, and survives critical testing. Transferring understanding does the trick, so can it be a dialogue's primary goal, after all?

This is a curious result. When epistemic luck is concerned, justification and understanding switch roles: seeking justification forces the Doctor to examine AI's performances in other cases (because true conclusions shouldn't be lucky), while seeking understanding allows her to focus on the actual patient (because understanding-related inferences *can* be lucky).

What do we conclude from all this? We started from the intuition that a diagnostic dialogue is about knowledge and justification rather than understanding. Issues concerning epistemic luck force us to rethink our approach and take a step back. Rather than exploring complex relationships between justification (knowledge) and understanding, we should further explore how they relate to the *truth*, i.e., to what extent, and in which circumstances, they contribute to accurately diagnosing a particular patient. The latter is the primary goal of the diagnostic Doctor-AI dialogue and all other (epistemic) abilities and states—justification, understanding, explanation, argument, clarification—are welcome as long as they promote it.

7. CONCLUSION

In this paper, we explored whether the model of explanatory dialogue—that focuses on the transfer of understanding—represents an optimal framework to analyze the Doctor-XAI interaction.

We showed that, typically, the Doctor is not after understanding, but justification. However, understanding and justification should ultimately serve the truth: what epistemic state the Doctor prefers—justified belief or understanding—depends on what promotes selecting an accurate diagnosis in particular circumstances. For instance, if the presence of epistemic luck would require rejecting an accurate diagnosis (because it is, in some sense, unjustified), the Doctor might privilege understanding over justification. At any rate, for studying the Doctor-XAI interaction, inquiry (Walton & Krabbe, 1995) is usually a more suitable normative framework than critical discussion (van Eemeren & Grootendorst, 2004). However, a critical discussion seems more suitable than an explanatory dialogue.

ACKNOWLEDGEMENTS: This paper was presented at the 10th International Conference on Argumentation (ISSA) (Leiden, July 4-7 2023). We thank the organizers, as well as the participants for their comments and suggestions. This publication was financially supported by the CHIST-ERA Programme via project CHIST-ERA/0002/2019.

REFERENCES

Cappelen, H. & J. Dever. (2021). *Making AI Explainable. Philosophical Foundations*. Oxford University Press.

Cawsey, A. (1992). *Explanation and interaction. The Computer Generation of Explanatory Dialogues*. The MIT Press: Cambridge, Massachusetts.

de Regt H (2015). Scientific understanding: truth or dare? *Synthese*, *192*, *12*, 3781-3797.

Douven, I. (2021). Abduction. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), URL = https://plato.stanford.edu/archives/sum2021/entries/abduction/

Dragulinescu, S. (2016). Inference to the best explanation and mechanisms in medicine. *Theoretical Medicine and Bioethics, 37*, 211-232, DOI 10.1007/s11017-016-9365-9

Dretske, F. I. (1972). Contrastive statements. *The Philosophical Review*, *81*, *4*, 411-437.
Dufour, M. (2017). Argument or Explanation: Who is to Decide? *Informal Logic*, *37*, *1*, 23-41.
Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, *132*, *1*, 33-42.
Faes, L., Sim, D., van Smeden, M., Held, U., Bossuyt, P., & L. Bachmann. (2022). Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins? *Front. Digit. Health*, Jan 26;4:833912. doi: 10.3389/fdgth.2022.833912. PMID: 35156082; PMCID: PMC8825497.
Gaszczyk, G. (2023). Helping Others to Understand: A Normative Account of the Speech Act of Explanation. *Topoi*, *42*, 285-396. https://doi.org/10.1007/s11245-022-09878-y
Goldman, A. (2003). An Epistemological Approach to Argumentation. *Informal Logic*, *23*, *1*, 51-63.
Govier, T. (2018). *Problems in Argument Analysis and Evaluation* (updated edition). Windsor Studies in
Argumentation (Volume 6): Windsor Ontario Canada.
Grimm, S. (2010). The goal of explanation. *Studies in History and Philosophy of Science*, *41*, 337-344.
Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*, *1*, 88-95. https://doi.org/10.2307/2183532
Hills, A. (2009). Moral testimony and moral epistemology. *Ethics 120*, *1*, 94-127.
Hills, A. (2016). Understanding why. *Noûs*, *49*, *2*, 661-688.
Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and H. Müller. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*, 9, 4, e1312.
doi: 10.1002/widm.1312. Epub 2019 Apr 2. PMID: 32089788; PMCID: PMC7017860.
Josephson, John R., & Susan G. Josephson. (1994). *Abductive Inference: Computation, Philosophy,\ Technology*. New York, Cambridge University Press.
Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. New York: Cambridge University Press. doi:10.1017/9781108164276.
Lipton, P. (2004). *Inference to the best explanation* (2nd edition). Routledge.
Marcum, J. A, (2008). *Humanizing Modern Medicine. An Introductory Philosophy of Medicine*. Springer Science + Business Media B.V.
Mayes, R. (2010). Argument-Explanation Complementarity and the Structure of Informal Reasoning. *Informal Logic 30*, *1*, 92-111.
McKeon, M. W. (2013). On the Rationale for Distinguishing Arguments from Explanations. *Argumentation*, *27*, 283-303.
Moore, D. J. (1995). *Participating in Explanatory Dialogues. Interpreting and Responding to Questions in Context*. Cambridge, Massachusetts: The MIT Press.
Nyrup, R. & D. Robinson. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical AI. *Ethics and Information Technology 24*:*13*. https://doi.org/10.1007/s10676-022\09632-3
Olmos, P. (2021). Metaphilosophy and Argument: The Case of the Justification of Abduction. *Informal Logic*, *41*, *2*, 131-164.
Pritchard, D. (2007). Anti-Luck Epistemology. *Synthese*, *158*, 277-98.
Rescher, N. (2007). *Dialectics. A Classical Approach to Inquiry*. Frankfurt: Ontos Verlag.
Scriven, M. (1972). The concept of comprehension: From semantics to software. In J. B. Carroll & R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge*. (pp. 31-39). Washington: W.
H. Winston & Sons.
Selbst, A. & Barocas, S. (2018). The intuitive appeal of explainable machine. *Fordham Law Review*, *87*, 1085-1139. Accessed 1 Aug 2019. https:// ir. lawnet. fordh am. edu/ flr/ vol87/ iss3/ 11.
Theunissen, M., and J. Browning. (2022). Putting explainable AI in context: institutional explanations for medical AI. *Ethics and Information Technology 24*, *23*, 1-10. https://doi.org/10.1007/s10676-022-09649-8.
van Eemeren, F.H., & R. Grootendorst. (2004). *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press.
van Laar, J.A., & Krabbe, E.C.W. (2013). The Burden of Criticism: Consequences of Taking a Critical Stance. *Argumentation 27*, *2*, 201-224.
Wagemans, Jean H. M. (2016). Argumentative Patterns for Justifying Scientific Explanations. *Argumentation*, *30*, 97-108.
Walton, D. (2004). A New Dialectical Theory of Explanation. *Philosophical Explorations*, *7*, *1*, 71-90.
Walton, D. (2005). *Abductive reasoning*. Tuscaloosa, Alabama: The University of Alabama Press.
Walton, D. (2007). The speech act of clarification in a dialogue model. *Studies in communication*

sciences, *7*, *2*, 165-197.
Walton, D. (2011). A dialogue system specification for explanation. *Synthese*, *182*, 349-374.
Walton, D., & E.C.W. Krabbe. (1995). *Commitment in dialogue. Basic concepts of interpersonal reasoning*. Albany: Suny Press.
Walton, D., C. Reed, & F. Macagno. (2008). *Argumentation schemes*. New York: Cambridge University Press.
Wright, L. (2002). Reasoning and Explaining. *Argumentation*, *16*, 33-46.
Yu, S., and Zenker, F. (2018). Peirce Knew Why Abduction Isn't IBE—A Scheme and Critical Questions for Abductive Argument. *Argumentation*, *32*, 569-587.