Minds and Machines (2024) 34:6
https://doi.org/10.1007/s11023-024-09657-1

# Towards a Benchmark for Scientific Understanding in Humans and Machines

**Kristian Gonzalez Barman**[1] · **Sascha Caron**[2,3] · **Tom Claassen**[4] · **Henk de Regt**[1]

Received: 25 May 2023 / Accepted: 22 February 2024 / Published online: 25 April 2024


**Abstract**
Scientific understanding is a fundamental goal of science. However, there is currently no good way to measure the scientific understanding of agents, whether these be humans or Artificial Intelligence systems. Without a clear benchmark, it is challenging to evaluate and compare different levels of scientific understanding. In this paper, we propose a framework to create a benchmark for scientific understanding, utilizing tools from philosophy of science. We adopt a behavioral conception of understanding, according to which genuine understanding should be recognized as an ability to perform certain tasks. We extend this notion of scientific understanding by considering a set of questions that gauge different levels of scientific understanding, covering information retrieval, the capability to arrange information to produce an explanation, and the ability to infer how things would be different under different circumstances. We suggest building a Scientific Understanding Benchmark (SUB), formed by a set of these tests, allowing for the evaluation and comparison of scientific understanding. Benchmarking plays a crucial role in establishing trust, ensuring quality control, and providing a basis for performance evaluation. By aligning machine and human scientific understanding we can improve their utility, ultimately advancing scientific understanding and helping to discover new insights within machines.

**Keywords** Benchmarking · Scientific understanding · Scientific Understanding Benchmark · Explanation · Counterfactual inference

## 1 Introduction

This paper presents a framework for measuring scientific understanding in agents, including humans, machine learning models, and model-augmented humans (Clark & Chalmers, 1998; Kuorikoski & Ylikoski, 2015). Current benchmarks in Machine

Extended author information available on the last page of the article

Learning measure a variety of capabilities (Thiyagalingam et al., 2022; Li & Zhan, 2022). For example, the Winograd Schema Challenge (WSC) (Levesque et al., 2012) and the General Language Understanding Evaluation (GLUE) (Wang et al., 2018) measure linguistic understanding, while BIGBench (Srivastava et al., 2022) measures proficiency at several tasks such as simple logic problems or guessing a chess move. However, despite their need and importance, there are currently no benchmarks that measure the degree of scientific understanding. To address this gap, we provide definitions of scientific understanding, a framework for how it can be measured, and we discuss potential use cases such as discovering new insights within machines.

The *main aim* of this paper is to provide a philosophical framework for benchmarking and measuring the scientific understanding of various agents[1], including Large Language Models (Vaswani et al., 2017; Brown et al., 2020) and Question Answering Machines (Allam & Haggag, 2012). Although our focus is on scientific understanding in the natural sciences (e.g., physics), we anticipate that our framework can be applied to other scientific disciplines as well. We break with the traditional view (Dellsén, 2020; Wilkenfeld, 2013; Searle, 1980; Johnson-Laird, 2010; Nersessian, 1992) by arguing that understanding should be conceptualized in terms of abilities rather than internal mechanics (Marcus, 2018; Chollet, 2017) or representations (Tamir & Shech, 2023; Wilkenfeld, 2013). Specifically, we contend that scientific understanding is a skill-based capability that relies on an agent's ability to perform specific actions, rather than a subjective mental state. This perspective separates the subjective 'feeling' of understanding from genuine understanding, indicating that psychological states are neither sufficient nor necessary to establish understanding (Rozenblit & Keil, 2002). Following De Regt (2017), we start from the idea that scientific understanding involves the ability to provide explanations within a theoretical framework that is intelligible to the agent, which includes the abilities to derive qualitative results, answer questions, solve problems properly, and extend knowledge to other domains or levels of abstraction. We argue that various degrees of scientific understanding can be measured by measuring different levels of ability, such as having access to relevant information, the ability to provide explanations, and the ability to establish counterfactual inferences. These different levels can be quantitatively evaluated using what-, why-, and w-questions (see Sect. 4).

Our framework enables creating specific tests, which can be used for benchmarking models, measuring student understanding, and evaluating teaching abilities, or training a machine learning model. We provide guidelines for researchers, including different testing interfaces. We then propose the creation of a benchmark for scientific understanding. This benchmark is an important first step towards assessing machine understanding, where aligning machine understanding with human understanding can help in research tasks, such as hypothesis creation and information retrieval and summarization.

It should be noted that the tests stemming from our framework differ significantly from the Turing Test (Turing, 1950; Oppy & Dowe, 2021). Their focus is not on

[1] By agents we mean actors who behave according to some set of (partially) internally generated rules. (partial) Autonomy, interactivity, and adaptability suffice for ascribing agency to an entity (intentionality or freedom are not required) (Jackson & Williams, 2021).

evaluating whether machines possess general intelligence but rather on measuring the degree of scientific understanding that any agent may have. The score one obtains in scientific understanding has no meaning as to whether an agent has AGI status or other abilities beyond scientific understanding. Similarly, passing the Turing Test does not necessarily mean an agent would score high on the scientific understanding test or vice versa.

We compare our framework to a recently proposed student-teacher interaction as a test of scientific understanding in machines (Krenn et al., 2022) and show how incorporating elements of our framework could improve this test and offer a quantifiable measure of transfer of understanding between agents. While their approach to testing focuses on evaluating new understanding, our framework is intended to work towards testing existing understanding as well as new understanding, where new understanding might involve increasing the understanding of phenomena (e.g., by providing deeper explanations) or discovering new phenomena. Finally, we discuss the abilities and limitations of our framework considering popular LLM implementations.

In sum, our paper makes the following contributions to the current debates on the role of AI and understanding in science:

- An analysis of scientific understanding as an ability, that should be measured in terms of behavioral competence (i.e., actions).
- A framework that provides a basis for developing tests to measure scientific understanding both in human and artificial agents. The framework can be used for benchmarking models, assessing student understanding, and training machine learning models.
- Guidelines for implementing tests to measure the scientific understanding of Large Language Models (LLMs) together with a call for scientific communities to systematically engage in benchmarking question-answering machines, to foster specific developments such as testing new scientific understanding (i.e., discovery of new insights unknown to humankind) within machines.
- A discussion on how to test whether a machine has transferred scientific understanding to another agent, building on a recently proposed account.

## 2 Scientific Understanding as an Ability: The Behavioral Conception of Understanding

Scientific understanding is traditionally viewed as an internal mental state or representation possessed by an agent, typically a human scientist (Baumberger et al., 2017; Grimm, 2021). This conception of understanding focuses on its subjective and internal aspects, from mental representations to the "feeling" of understanding, rather than the observable aspects of the agent's abilities and actions. Philosophers who criticize the idea that machines might be capable of scientific understanding often base their criticisms on this 'internalist' view of understanding. Floridi (2023), for example, supports his skepticism about the capacity of Large Language Models (LLMs) to achieve any degree of understanding with the claim that they do not

reason or resemble the cognitive processes present in animal or human brains. Following this line of thought, critics of LLMs often argue that agents must have the same underlying mechanisms as humans to understand. We submit, however, that this argument does not hold water, as it is unclear why understanding could not be realizable through different mechanisms. Moreover, human understanding is rarely assessed by means of inspecting underlying mechanisms, but rather via observation of (the results of) behavioral actions, such as answering questions in an exam or performing certain exercises.

We propose a re-evaluation of scientific understanding, arguing that it can and should be assessed on the basis of an agent's ability to perform certain tasks, rather than on the underlying mechanisms involved in those tasks. This implies that artificial agents, including Large Language Models (LLMs), should not be dismissed out of hand as being incapable of scientific understanding, simply because they allegedly "guess the next word" or are "stochastic parrots" (Bender et al., 2021).

The traditional conception of understanding is rooted in a supposed analogy between understanding and knowledge, where knowledge is taken to be justified true belief. Understanding is then assumed to be either a specific type of knowledge (a "species of knowledge", as Grimm, 2016 claims) or a kind of belief that is in some sense analogous to knowledge (see Baumberger et al., 2017, for an overview). However, this conception has been challenged in recent work by epistemologists and philosophers of science (Elgin, 2017; Potochnik, 2017; De Regt, 2017). Thus, De Regt (2023) observes that there are at least three problems for the thesis that understanding resembles knowledge. First, while knowledge presupposes truth (the so-called factivity requirement), understanding does not seem to be factive. The reason is that if understanding were factive, it cannot be had with idealized models and false theories, which would imply that much of past and present science fails to yield understanding. Second, criteria for scientific understanding vary strongly with the historical and disciplinary context, a fact that is hard to reconcile with the conception of understanding as a species of knowledge. Third, it seems intuitively clear that higher forms of scientific understanding require something more than mere knowledge: merely *knowing facts* about a particular phenomenon does not yet amount to understanding *why* that phenomenon occurs. Taken together, these three points lead to the suggestion that the crucial ingredient of understanding is not some kind of mental representation but something else that can in fact be directly observed: the skill, or ability, to use relevant knowledge. In particular, as we will argue below in Sect. 3, understanding involves the skill to construct explanations and use them in the right way. Whether or not (or to which degree) an agent possesses such skills can be observed and evaluated via testing procedures.[2] These considerations suggest that scientific understanding is, or at least involves, an ability that can and should be assessed via observation of an agent's behavior. The assumption that understanding requires specific mental representations, internal architecture, consciousness, or other similar factors, appears

[2] To be sure, knowledge possession can also be tested, but note that this never involves direct observation of the mental representation (belief) that knowledge allegedly consists in, but rather observation of an ability, for example the ability to answer questions correctly.

to be unsupported and indeed unfruitful when it comes to measuring the degree of understanding in an agent.

We maintain that the evaluation of understanding in any agent, including artificial ones like LLMs, should follow the same principles used to assess human scientific understanding — that is, it should be based on their abilities to perform relevant tasks. Incidentally, we are not the only ones who relate understanding to an ability (see Krenn et al., 2022; Tamir & Shech, 2023). For example, Tamir and Shech (2023) have argued that practical abilities (such as reliable and robust task performance) can be seen as key factors indicative of understanding in the context of deep learning. While we think this is a good start, we argue that a more comprehensive and rigorous evaluation of understanding as an ability is needed. Below, in Sects. 3–5, we will outline how this can be achieved.

A behavioral conception of scientific understanding, that defines understanding in terms of abilities to perform certain tasks, has several advantages. First of all, it is a deflationary approach to understanding that requires less ontological baggage than the traditional view. Moreover, and more importantly, it aligns better with how we think about and apply the notion of understanding in many real-life contexts, for example, when gauging other people's understanding. The behavioral conception emphasizes the epistemic and practical reliability of knowledge, highlighting its effectiveness in achieving specific objectives. This approach avoids the enigmatic nature of internal mental states and enables a straightforward metric, aligning with the trajectory of scientific progress, which often hinges on the development of practical applications and technologies which are publicly accessible and accountable.

Given that we have argued that understanding should be interpreted in terms of capabilities and behaviors, the question now arises as to which capabilities matter for scientific understanding. Here we adopt an 'empirical' approach: look at scientific practice for guidance on which abilities are exercised. Since scientific understanding is typically interpreted as explanatory understanding, we argue that tests for understanding should target the most common forms of abilities related to the construction and use of scientific explanations. These include, first of all, the ability to utilize the right ingredients for explanation formation (which we argue can be measured through what-questions, as explained in next section); next, the ability to build explanations (which we argue can be measured with why-questions); and finally, being able to use these explanations to generate counterfactual inferences (which we argue can be measured with w-questions).

In the next section we elaborate our framework for scientific understanding, after which, in Sects. 4 and 5, we detail a possible way to apply said framework towards creating tests.

## 3 A Framework for Scientific Understanding

Our starting point is De Regt's (2017) account of scientific understanding, on which understanding a phenomenon boils down to having an adequate explanation of the phenomenon within the right theoretical scaffolding. The formal criterion is the following (2017, p.92) (Criterion for Understanding a Phenomenon):

> **CUP:** A phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency.

Note that the use of a biconditional indicates that this is a necessary and sufficient condition. The explanation must be based on an intelligible theory. A test for intelligibility can described by the following criterion (2017, p.102):

> **$CIT_1$:** A scientific theory *T* (in one or more of its representations) is intelligible for scientists (in context *C*) if they can recognize qualitatively characteristic consequences of *T* without performing exact calculations.

In this case, the condition is sufficient but not necessary. This is also why there is a subscript 1 in **$CIT_1$**, since there might be other criteria for intelligibility. **$CIT_1$** holds primarily for theories with a mathematical formulation, such as in physics. For other types of theories, other conditions might hold. The key aspect of this condition is the ability to derive (qualitative) consequences.

To elaborate this conception of scientific understanding, we modify the definition by shifting the focus from the phenomenon being understood to the conditions required for an agent to understand. We develop these conditions into having access to information, having explanatory abilities (since they might not coincide), and reformulate the ability to recognize qualitatively characteristic consequences in terms of counterfactual inferences. Moreover, we refine the definition by emphasizing the importance of measuring understanding instead of relying on strict necessary and sufficient conditions. Doing so acknowledges that understanding is not a binary affair but rather a matter of degree. A graded notion of understanding allows for the recognition of varying levels of capabilities and has been argued for across the literature (Baumberger, 2019; Kelp, 2015). This approach enables a more nuanced evaluation of agents' progression, firstly because it is more granular, secondly, because it enables to draw comparisons between agents.

These extensions result in the following **general framework for an agent's scientific understanding** (**A**gent-**U**nderstands-**P**henomenon):

> **AUP:** The degree to which agent A scientifically understands phenomenon P can be determined by assessing the extent to which (i) A has a sufficiently complete representation of P; (ii) A can generate internally consistent and empirically adequate explanations of P; (iii) A can establish a broad range of relevant, correct counterfactual inferences regarding P.

**AUP** is our framework for establishing the degree of scientific understanding of a phenomenon (by an agent). This framework can be instantiated in different ways (e.g., there might be several ways of establishing whether A has a sufficiently complete representation of P). One implementation would be **$AUP_1$**:

> **$AUP_1$**(i-iii) can be measured, given a certain context (series of prompts) via what-, why, and w-questions respectively.

These questions are prompt specific, where the context (i.e.,. initial prompting to provide necessary information) and the ordering of questions makes a difference. This feature makes the application $\mathbf{AUP_I}$ dynamic.

The first level ('i') of **AUP** requires having access to sufficient relevant information about P. This access involves the capacity to retrieve information from relevant sources (such as memories, encodings/embeddings, databases, or the internet). We argue that this can be measured by the ability to provide correct answers to 'what-questions' (see Sect. 4.1).

The second level ('ii') refers to the capability of arranging information to produce an explanation of P. The ability to generate a well-constructed explanation surpasses simple information retrieval, requiring a deeper level of understanding (Woodward & Ross, 2021). We argue that the ability to provide explanations can be evaluated through answers to why-questions (see Sect. 4.2).

The third level ('iii') is the ability to infer how P would have been (or would be) different under different circumstances; namely, the ability to draw counterfactual inferences. This ability requires being able to properly *use* a (good) explanation (Woodward, 2003; Hitchcock & Woodward, 2003; Weslake, 2010). We argue that this ability can be measured via answers to w-questions (see Sect. 4.3). Answers to w-questions require more than simply having an explanation, they require having a good explanation and knowing how to use it (e.g., knowing when the explanation is applicable, what the boundary conditions are, etc.). W-questions assess competency at establishing counterfactual inferences concerning a phenomenon and can be linked to an agents' breadth and depth of understanding (Kuorikoski & Ylikoski, 2015).

These three levels capture a broad spectrum of capabilities, but more importantly, enable pinpointing where an agent stands in their journey towards greater understanding. It might be objected[3] that an ability to answer w-questions is sufficient for understanding. However, if we only accept answers to w-questions as proxies for understanding, we might lose important information. While if one is able to answer w-questions, one should also be able to answer why-questions, the reverse does not follow: one might be able to answer certain why-questions, or certain what-questions, and not be capable of answering counterfactual questions. We argue that in such cases there may still be understanding (albeit in a less sophisticated form). These forms of understanding become particularly important when trying to benchmark how close certain agents might be to reaching counterfactual understanding. If we simply ignore the ability to answer said questions, all we could say is that any agent that cannot reason counterfactually has no understanding.

## 4 Test Questions

### 4.1 What-questions

What-questions ask for descriptive knowledge about an object or phenomenon (Belnap & Steel, 1976; Cross & Roelofsen, 2022). Answering such questions requires

[3] We thank an anonymous reviewer for alerting us to this objection.

having access to information, whether from memory or external sources (books, servers, etc.). What-questions can ask for values, dimensions, or names, among other things. For example, what is the charge of the electron? The ability to answer what-questions correctly is a necessary but not a sufficient condition for higher levels of understanding, as it only involves the retrieval of information and not the ability to use said information for higher-level tasks (see section above).

### 4.2 Why-Questions (Explanation-Seeking Questions)

Answering why-questions (Cross, 1991; Bromberger, 1966; Hempel & Oppenheim, 1948) which inquire about facts or phenomena requires providing an explanation. It is for this reason that answering them correctly is highly indicative of scientific understanding. Why-questions can be divided into (at least) three types:

1. Questions of singular facts: 'Why is it the case that A?' ('Why is charge conserved?').
2. Contrastive questions (van Fraassen, 1980): 'Why A rather than B' ('Why did Patient A rather than Patient B get better with treatment T?', 'Why did Patient P get better using treatment A rather than treatment B?').
3. Resemblance questions(Weber & Lefevere, 2017): 'Why do A and B share C' ('Why do both hedgehogs and bears hibernate?').

Answering why-questions requires articulating information in a way that is sensitive to context and explanatory aims (van Fraassen, 1980). Using a variety of why-questions with answers not easily found (e.g., by choosing different foils or contrast classes in the case of contrastive explanations) can help ensure an explanatory ability that is not simply due to memorization or accessing the internet.

### 4.3 W-questions (Counterfactual Inferences)

W-questions refer to what-if-things-had-been-different questions (Woodward, 2003) and what-would-happen-if questions (Weber et al., 2019). These questions explore alternative scenarios and potential outcomes based on a hypothetical change in circumstances. Answers to what-if-things-had-been-different questions enable us to see what the outcome of some state of affairs would have been if initial conditions had been different. Answers to what-would-happen-if questions can be thought of as a prediction that involves some sort of manipulation or intervention on a system (Weber et al., 2019).

Answering these two types of questions can be thought of as backward-looking and forward-looking counterfactual inferences (Barman & van Eck, 2021). In both cases, answering these questions involves postulating hypothetical scenarios about what would occur under a specific set of circumstances. It is this feature that we are interested in, since the ability to adequately derive these scenarios requires understanding. Similarly, there is a strong link between the quality of an explanation and the counterfactual inferences it affords (Woodward, 2003; Ylikoski & Kuorikoski, 2010; Barman, 2022). We therefore contend that the range of counterfactual infer-

ences an agent can articulate is strongly related to the level of understanding (Kuorikoski & Ylikoski, 2015). Counterfactual inferences can be more or less general, where one can distinguish between parameterized and exogenous variable counterfactual inferences (Pearl, 2009; Halpern, 2016). Parameterized counterfactual inferences involve changing specific variables within a model, such as "What would happen to the period of this spring-mass system if we changed the spring constant?". Exogenous variable counterfactual inferences involve changing external variables that impact the system, such as "What would happen to this spring-mass system if the spring breaks?".

## 5 From a General Framework to Specific Tests

In this section we discuss how to operationalize the framework described in the previous section into concrete tests to measure scientific understanding. Understanding can be of a concrete phenomenon (e.g., a pendulum of 5 m length, 2 kg weight, etc.) or of a general phenomenon (e.g., pendulums in general). Tests can be devised for both specific and general phenomena, and the level of generality can be increased by asking higher-level w-questions, such as those related to changing exogenous variables.

### 5.1 How to Score an Agent?

The level of scientific understanding of an agent can be thought of as a gradient between complete lack of understanding to an ever-increasing level (See Fig. 1). The agent's score would depend on the number of correct answers, with varying weights assigned to different questions. This test can provide a specific score for the scientific understanding of an agent or compare two agents. We can establish different thresholds depending on the context.

Each individual test should contain a sufficiently diverse and representative set of questions that can capture enough details of the properties, attributes, and elements of the phenomenon. This implies that question generation should be conducted by the experts of each community. We provide some guidelines for this below. In some cases, it could be possible to train language models to produce questions as well (Perez et al., 2022; Du et al., 2017; Rao & Daumé III, 2018).

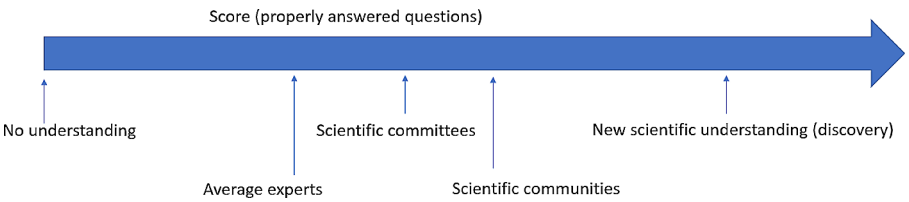


**Fig. 1** Scientific understanding can be categorized into various levels based on the number of questions answered. An agent possessing the ability to answer all the questions posed by scientific communities, including those for which we do not yet have answers, indicates a higher level of (new) scientific understanding

### 5.2 Guidelines for Testing

There is a need for guidelines to establish a standardized and reliable approach to testing which ensures accurate and consistent results. In developing a comprehensive and reliable test for AI agents, it is crucial to define the test scope and purpose, ensuring it is tailored to the agent and includes a variety of difficulty levels. To achieve consistent, repeatable scoring, diverse question formats should be employed, with multiple testing instances conducted for robust evaluation. Crafting comprehensive, varied, and representative questions is essential, using concise and unambiguous language to prevent confusion. To maintain test integrity, limit answer accessibility on the internet and other public repositories. Finally, centralized storage of tests for easy review, enabling them to serve as part of the SUB benchmark that we will introduce below. Additional guidelines for good testing and evaluation can also be implemented (Mintzes et al., 2005; Schleicher, 1999; Franzen, 2010; Brookhart, 2013).

It is important to note that this test should be conducted through an interface, which may need to be tailored for certain agents. Traditional testing methods such as multiple-choice tests can be used for humans and model-augmented humans, while an interface that allows for context encoding and some form of chat-like interface (e.g., ChatGPT[4]) is needed for LLMs.

### 5.3 The Scientific Understanding Benchmark (SUB)

After describing our framework for testing understanding, we would in addition like to propose two things that, while not the main objective of this paper, are nonetheless fundamental for its proper implementation. First, a call to communities to create tests for scientific understanding to benchmark different AI models. Benchmarking plays a crucial role in establishing trust in the reliability of models, ensuring quality control, and providing a basis for performance evaluation. Given the current situation in AI it is thereby of high societal relevance. Second, the bringing together of tests developed by different communities into a broad benchmark, which we call the Scientific Understanding Benchmark (**SUB**). This should be an open project supervised by an independent community of experts that, among other criteria, sets high standards for scientific correctness.

We firmly believe that the **SUB** will have a positive impact on the usefulness, confidence, and controllability of AI in scientific research and expect it to advance scientific understanding, facilitate stakeholder alignment, and enable new discoveries.

## 6 Scientific Understanding Transfer

Krenn et al. (2022) closely follow an earlier version of de Regt's (2017) account, namely, de Regt and Dieks (2005). Based on **CIT** (see the section on Scientific Understanding) they formulate a parallel condition replacing the scientist(s) with an AI. Subsequently, they add an additional condition, according to which 'An AI gained

[4] https://openai.com/blog/chatgpt.

scientific understanding if it can transfer its understanding to a human expert' (2022, p. 767). They then combine these two conditions into a test, which they describe as follows (Ibidem):

> A human (the student) interacts with a teacher, either a human or an artificial scientist. The teacher's goal is to explain a scientific theory and its qualitative, characteristic consequences to the student. Another human (the referee) tests both the student and the teacher independently. If the referee cannot distinguish between the qualities of their non-trivial explanations in various contexts, we argue that the teacher has scientific understanding.

While promising, this approach may have a few issues. First, it equates teaching abilities with understanding, which is problematic if the student is simply a bad learner (despite the teacher's understanding). Second, the referee determines understanding by comparing the qualities of explanations. If both teacher and student lack understanding (whether because they simply lack explanations or because their explanations are incorrect), their explanations may be indistinguishable (by being equally wrong). According to the test, we should conclude the teacher has understanding. Third, the test is difficult to implement in practice due to vague parameters, such as the referee's inability to distinguish between non-trivial explanations in different contexts. The quality of explanations depends on explanatory aims and the variety of contexts is unclear, leading to different results depending on the chosen referee.

Despite its limitations, we think Krenn et al.'s test is valuable, and we propose a reformulation of it using our framework. Instead of relying on a referee, we propose to measure the student's score before and after interacting with a teacher to demonstrate an increase in scientific understanding (by the student). Additionally, we can test the teacher's understanding separately to distinguish between the teacher's own understanding and their ability to transfer that understanding to the student. We then suggest the reformulated test for scientific understanding transfer:

> The student takes an initial test, interacts with the teacher, and then takes a second test. While the second test should cover the same material or aspects, it should contain different questions to ensure the validity of the test. The extent to which the student's score increases on the second test is an indication of the teacher's ability to effectively phenomenon P to the student.

We view this reformulation as an improvement because it enables measuring the increase of scientific understanding in agents. This becomes particularly important when AI has developed new knowledge that needs to be conveyed to humans who lack that understanding. The reformulated test can be helpful in important cases where AI has developed new knowledge that needs to be conveyed to humans.

## 7 Applications, Limitations, and New Scientific Understanding

The usefulness of AI models, such as Large Language Models, in scientific contexts like hypothesis generation and information retrieval relies on aligning their scientific understanding with humans. The proposed framework for assessing and directing scientific understanding in AI agents has the potential to enhance the usefulness of AI models in scientific contexts. For instance, it can compare the performance of different AI agents in answering questions, as well as highlighting their strengths and limitations. Additionally, it can also aid in educational programs. By helping select relevant AI tools, this framework could be a valuable resource for students, serving as a pedagogical aid. Some AI models are already capable of performing above the level of a college student who has completed one semester of physics (West, 2023), highlighting the potential of these models as a valuable resource for students and researchers alike in the near future.

However, there are still open questions and challenges that need to be addressed. Establishing a threshold to determine sufficient understanding for an agent can prove to be complex, particularly when there may not be a consensus on the appropriate criteria. Similarly, in some testing modalities, experts might not always be available to check what the correct answer is, and for some questions we simply do not yet know the answer. However, this could open the possibility for new avenues of research, as asking these questions to QAMs might in some cases provide interesting answers that can trigger new avenues for research and stimulate hypothesis generation (Bubeck et al., 2023). In such a case, if the agent is capable of answering questions for which humans do not have an answer yet, it may possess new scientific understanding.

To evaluate this *new scientific understanding*, a community can define w-questions whose answers are not yet known, similar to conjectures in mathematics that can be expected to be solved soon (Ganesalingam & Gowers, 2017), where AI can tentatively provide unverified answers[5]. These lists of w-questions can measure progress in gaining new scientific knowledge and test forms of new scientific understanding. It is important to determine whether current LLMs/QAMs have new scientific understanding by asking questions where we might not know the answer. One can then employ our framework to act as a retroactive new scientific understanding test once these discoveries are confirmed or denied. This approach can encourage discovery and stimulate further research.

Finally, choosing the right prompt (Reynolds & McDonell, 2021)(e.g., what context needs to be provided for each question) and evaluating vague outputs can be challenging. The Reinforcement Learning (Sutton & Barto, 2018) used to fine tune certain models (optimizing for user experience) may steer in the direction of vague or incorrect answers which are detrimental to scientific research (Perez et al., 2022). Addressing this issue may require AI models specifically designed for scientific aims, which could involve adjusting the training set (Taylor et al., 2022) or prioritizing scientific understanding during training or fine tuning. Creating AI models tailored to scientific understanding has the potential to transform how we approach scientific exploration.

[5] https://ai.facebook.com/blog/ai-math-theorem-proving/.

## 8 Conclusion

This paper presents a philosophical framework for creating tests that assess agents' scientific understanding. It provides discussions and guidelines for communities to create their own scientific understanding tests, stressing their importance. The potential impact of this framework is multifold, as it can enhance the usefulness of AI, assess possible new scientific understanding encoded in machines, and aid in educational programs.

Future research directions include refining the methodology for creating tests and the concrete elaboration of tests which will form part of the benchmark for scientific understanding. Ultimately, scientific understanding tests are necessary to analyze, control, and harness the potential of AI in the context of scientific research.

**Acknowledgements** We would like to express our gratitude to the participants of the International Association for Computing and Philosophy (IACAP), and the participants of the European Philosophy of Science Association 2023 conference (EPSA23), where this paper was presented. Additionally, we would like to thank two anonymous reviewers whose constructive feedback significantly enhanced the quality of this paper.



## References

Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Barman, K. G. (2022). *Procedure for assessing the quality of explanations in failure analysis* (p. 36). AI EDAM.

Barman, K. G., & van Eck, D. (2021). IBE in engineering science-the case of malfunction explanation. *European Journal for Philosophy of Science*, *11*, 1–19.

Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science*, *50*(3), 367–388.

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In Explaining understanding: new perspectives from epistemology and philosophy of science. Eds. Grimm, S. R., Baumberger, C., and Ammon S. Routledge (pp.1–34).

Belnap, N. D., & Steel, T. (1976). B. The logic of questions and answers.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 列. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610–623) (2021, March).

Bromberger, S. (1966). Why-questions. In R. G. Colodny (Ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy* (pp. 86–111). University of Pittsburgh.

Brookhart, S. M. (2013). *How to create and use rubrics for formative Assessment and Grading*. ASCD.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.
Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint *arXiv:2303.12712*.
Chollet, F. (2017). The limitations of deep learning. *Deep Learning with Python*.
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.
Cross, C. B. (1991). Explanation and the theory of questions. *Erkenntnis*, *34*(2), 237–260.
Cross, C., Roelofsen, F., & Questions The Stanford Encyclopedia of Philosophy, E. N. Zalta (Ed.), (Summer 2022 Edition).
De Regt (2017). *H. W. understanding scientific understanding*. Oxford University Press.
De Regt, H. W. (2023). Can scientific understanding be reduced to knowledge? In Scientific Understanding and Representation: Modeling in the Physical Sciences. Eds. Lawler, I., Khalifa, K., and Shech, E. Routledge (pp. 17–32).
De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 137–170.
Dellsén, F. (2020). Beyond explanation: Understanding as dependency modelling. *The British Journal for the Philosophy of Science*.
Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *arXiv Preprint arXiv*:170500106.
Elgin, C. (2017). *Z. True enough*. MIT Press.
Floridi, L. (2023). AI as Agency without Intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15.
Franzen, M. (2010). Assessing student understanding in Science. *Science and Children*, *47*(9), 79.
Ganesalingam, M., & Gowers, W. T. (2017). A fully automatic theorem prover with human-style output. *Journal of Automated Reasoning*, *58*, 253–291.
Grimm, S. R. (2016). Is understanding a species of knowledge? *The British Journal for the Philosophy of Science*, *57*, 515–535.
Grimm, S. R., & Understanding The Stanford Encyclopedia of Philosophy. Edward N. Zalta (Ed.) (Summer 2021 Edition).
Halpern, J. Y. (2016). *Actual causality*. MIT Press.
Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.
Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.
Jackson, R. B., & Williams, T. (2021). A theory of social agency for human-robot interaction. *Frontiers in Robotics and AI, 8*, 687726.
Johnson-Laird, P. N. (2010). Mental models and human reasoning. Proceedings of the National Academy of Sciences, 107(43), 18243–18250.
Kelp, C. (2015). Understanding phenomena. *Synthese*, *192*(12), 3799–3816.
Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., Yao, Z., & Aspuru-Guzik, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, *4*(12), 761–769.
Kuorikoski, J., & Ylikoski, P. (2015). External representations and scientific understanding. *Synthese*, *192*, 3817–3837.
Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge. KR, 13th (2012).
Li, Y., Zhan, J., & SAIBench (2022). Benchmarking AI for science. *BenchCouncil Transactions on Benchmarks Standards and Evaluations*, *2*(2), 100063.
Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.
Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (Eds.). (2005). *Assessing Science understanding: A human constructivist view*. Academic.
Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. *Cognitive Models of Science*, *15*, 3–44.
Oppy, G., & Dowe, D. The Turing Test. The Stanford Encyclopedia of Philosophy, Edward N. Zalta (Ed.) (Winter 2021 Edition).
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys Vol*, *3*, 96–146.
Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., & Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. arXiv preprin*t arXiv:2212.09251*.
Potochnik, A. (2017). *Idealization and the aims of science*. The University of Chicago.

Rao, S., & DauméIII, H. (2018). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv Preprint arXiv*:180504655.
Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–7).
Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.
Schleicher, A. (1999). *Measuring Student Knowledge and skills: A New Framework for Assessment*. Organisation for Economic Co-Operation and Development.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.
Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.
Tamir, M., & Shech, E. (2023). Machine understanding and deep learning representation. *Synthese*, *201*(2), 51.
Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., & Stojnic, R. (2022). Galactica: A large language model for science. *arXiv Preprint arXiv*:221109085.
Thiyagalingam, J., Shankar, M., Fox, G., & Hey, T. (2022). Scientific machine learning benchmarks. *Nature Reviews Physics*, *4*(6), 413–420.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*(236), 433–460.
Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv Preprint* arXiv:1804.07461(2018).
Weber, E., & Lefevere, M. (2017). Unification, the answer to resemblance questions. *Synthese*, *194*, 3501–3521.
Weber, E., van Eck, D., & Mennes, J. (2019). On the structure and epistemic value of function ascriptions in biology and engineering sciences. *Foundations of Science*, *24*, 559–581.
Weslake, B. (2010). Explanatory depth. *Philosophy of Science*, *77*(2), 273–294.
West, C. G. (2023). AI and the FCI: Can ChatGPT Project an Understanding of Introductory Physics? arXiv preprint arXiv:2303.01067.
Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, *190*, 997–1016.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
Woodward, J., Ross, L., & Scientific Explanation The Stanford Encyclopedia of Philosophy. Edward N. Zalta (Ed.) (Summer 2021 Edition).
Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*, 201–219.

## Authors and Affiliations

**Kristian Gonzalez Barman[1] · Sascha Caron[2,3] · Tom Claassen[4] · Henk de Regt[1]**

✉ Henk de Regt
henk.deregt@ru.nl

Kristian Gonzalez Barman
KristianCampbell.GonzalezBarman@UGent.be

Sascha Caron
scaron@nikhef.nl

Tom Claassen
tomc@cs.ru.nl

[1] Institute for Science in Society, Faculty of Science, Radboud University the Netherlands, Nijmegen, the Netherlands

[2] High Energy Physics, Faculty of Science, Radboud University the Netherlands, Nijmegen, the Netherlands

[3] Nikhef, Science Park 105, Amsterdam 1098 XG, the Netherlands

[4] Institute for Computing and Information Sciences, Faculty of Science, Radboud University, Nijmegen, the Netherlands