“Foundation models for research A matter of trust?”

Bruynseels, Koen; Asveld, Lotte; van den Hoven, Jeroen

DOI
10.1016/j.ailsci.2025.100126

Publication date
2025

Document Version
Final published version

Published in
Artificial Intelligence in the Life Sciences

Citation (APA)
Bruynseels, K., Asveld, L., & van den Hoven, J. (2025). “Foundation models for research: A matter of trust?”. *Artificial Intelligence in the Life Sciences*, *7*, Article 100126. https://doi.org/10.1016/j.ailsci.2025.100126

Artificial Intelligence in the Life Sciences 7 (2025) 100126

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

“Foundation models for research: A matter of trust?”

Koen Bruynseels [a,*], Lotte Asveld [b], Jeroen van den Hoven [a,b]

[a] TU Delft, Digital Ethics Centre, Jaffalaan 5 2628BX Delft, The Netherlands
[b] TU Delft, Ethics & Philosophy of Technology, Jaffalaan 5 2628BX Delft, The Netherlands

ARTICLE INFO

Keywords:
Foundation models
Large language models
AI agents
Epistemic
Trust
Reliance

ABSTRACT

Science would not be possible without trust among experts, trust of the public in experts, and reliance on scientific instruments and methods. The rapid adoption of scientific foundation models and their use in AI agents is changing scientific practices and thereby impacting this epistemic fabric which hinges on trust and reliance. Foundation models are machine learning models that are trained on large bodies of data and can be applied to a multitude of tasks. Their application in science raises the question of whether scientific foundation models can be relied upon as a research tool and to what extent, or even be trusted as if they were research partners.

Conceptual clarification of the notions of trust and reliance in science is pivotal in the face of foundation models. Trust and reliance form the glue for the increasingly distributed epistemic labour within contemporary technoscientific systems. We build on two concepts of trust in science, namely trust in science as shared values, and trust in science based on commitments to processes that provide objective claims. We analyse whether scientific foundation models are research tools to which the concept of reliance applies, or research partners that can be trustworthy or not. We consider these foundation models within their socio-technical contexts.

Allocation of trust should be reserved for human agents and the organizations they operate in. Reliance applies to foundation models and artificial intelligence agents. This distinction is important to unambiguously allocate responsibility, which is crucial in maintaining the fabric of trust that underpins science.

1. Introduction: the impact of foundation models on notions of epistemic trust and reliance

Trust is central to the scientific enterprise, as it is impossible for a person to acquire all knowledge firsthand. Members of the public need to assess whether to put trust in the claims of scientists. And scientists have to rely on their fellow scientists to build their knowledge claims. Trust is the glue of teamwork and makes the division of labour possible [1].

Research activities increasingly include Artificial Intelligence (AI) models that are put to specific scientific tasks. These models, such as Large Language Models (LLMs) and other forms of Generative Artificial Intelligence (GenAI), can capture patterns in large bodies of information. Since research is an increasingly data-intensive activity, these models are successfully applied for scientific inference and discovery. The increased use of scientific foundation models in research raises the question of how these models should be incorporated into the process of knowledge generation and discovery.

An increasing number of projects aim at the development of foundation models that capture data domains central to a scientific field. For instance, in computational biology, models like AlphaFold [2], ESMFold [3] and RosettaFold [4] have transformed the field of protein structure prediction. These models allowed for the discovery of novel drug candidates, and even the generation of novel proteins that are not found in nature. These foundation models are trained on data gathered in public repositories, like the PDB protein data bank [5] that contains protein structures as inferred over decades of experimental work. Similarly, in material science, DeepMind’s Graph Networks for Materials Exploration (GNoME) [6] was trained with data on crystal structures and related properties, freely accessible via the Materials Project [7]. This foundation model enabled the discovery of new materials, some of which can lead to the development of transformative technologies ranging from superconductors to next-generation batteries. In the research field of microbiology, explainable deep learning led to the discovery of new candidate antibiotics. The model predicted nearly a million new antibiotics, with many of the hundred tested candidates showing *in vitro* efficacy against pathogens [8].

These examples are indicative of the enormous potential that

* Corresponding author.
E-mail addresses: K.R.C.Bruynseels@tudelft.nl (K. Bruynseels), L.Asveld@tudelft.nl (L. Asveld), M.J.vandenHoven@tudelft.nl (J. van den Hoven).

https://doi.org/10.1016/j.ailsci.2025.100126
Received 29 November 2024; Received in revised form 2 February 2025; Accepted 3 February 2025
Available online 6 February 2025

foundation models hold for scientific discovery. As the impact of these models grows, so too does the need to assess their reliability and role within the scientific process. The integration of foundation models into scientific research is not without challenges. Concerns about the trustworthiness of these models have emerged, particularly in relation to issues such as the hallucinations generated by Artificial Intelligence (AI), where models produce incorrect or misleading information. An early example was Meta's Galactica, which was trained on a large body of scientific literature, and encountered difficulties in producing reliable scientific claims [9]. Significant progress has though been made in tackling scientific queries [10], as exemplified in the OpenAI o1 models [11], and in the DeepSeek-R1 model [12]. One strategy is to mimic human reasoning processes by having a LLM split up complex questions into logical steps, developing an argumentation that leads from the premises to the conclusion. These Chain of Thought (CoT) prompting techniques increase the accuracy in reasoning-intensive tasks. Another approach, "Retrieval Augmented Generation" (RAG), enhances the reliability of the output of the LLM by augmenting the query with information from dedicated knowledge repositories.

Since LLMs lack intentionality, they are indifferent to the truth of their outputs. Care for the truth thus needs to be explicitly and successfully included in the design of these systems. The output of LLMs like ChatGPT was therefore even framed as "bullshit", in the sense explored by Harry Frankfurt as a lack of concern for truth [13]. This is further complicated by the limits to the explainability of the outputs, which is due to the inherent opaqueness of LLMs. The latent space of a deep learning model provides a compressed representation of the features that the model learned from the datasets it was trained on. Explainability can be pursued via an analysis of this latent space in the machine learning model. The ability to derive a higher level logic within the model that allows for explaining relations between input and output though has limitations, and this might remain so in the future [14]. Current research in Explainable Artificial Intelligence (XAI) focuses on local rather than global interpretability. Another source of opaqueness is the proprietary context in which foundation models are often developed. This double 'black box' problem presents significant epistemic challenges, particularly regarding explainability and transparency. For example, Google DeepMind did not make the source code available when releasing the AlphaFold3 algorithm for protein structure prediction. This raised concerns within the academic community about dependency on big tech, and the inability to fully and freely access the model and explore its scientific potential. Efforts to develop alternative models that are more transparent were initiated as a result [15], and the code of AlphaFold3 was released six months after the release [16].

The exchange between Elon Musk and Yann LeCun on this topic highlights the widely divergent views on whether foundation models in the scientific domain can be genuinely trusted or relied upon [17]. The focal point of this discussion was the xAI project, which aims at creating an AI dedicated to "understanding the universe through a maximally rigorous pursuit of truth, unburdened by concerns of popularity or political correctness". Such an ambitious epistemic goal, and the big tech context in which the foundation model will be developed, raises questions about the role and impact of AI on scientific inquiry. Fundamentally, it raises the question about the shifting nature of the scientific enterprise itself.

This paper focuses on the question of whether it conceptually makes sense to consider foundation models as research tools that can be relied on, or rather as research partners that can be trusted. We will analyze the validity of the concepts of reliance and epistemic trust in the face of foundation models as an emerging epistemic technology. Both trust and reliance are required during the development of technological artefacts, as well as in scientific practices. Without some level of trust in experts and reliance on scientific tools, it is impossible to gain knowledge. How these concepts are understood thus impacts the very fabric of the techno-scientific enterprise. Different concepts of trust and reliance will lead to different socio-technical and institutional designs and different scientific practices. Trust and reliance though are indeterminate concepts to a certain degree. In this paper, we will therefore pursue a "conceptual engineering" [18,19] of the concepts of trust and reliance in epistemic settings. To achieve this, we will explore two distinct notions of epistemic trust in science, and contrast them with epistemic reliance. Since both trust and reliance are relational concepts, we consider them in relation to their potential recipients. Firstly, we will examine whether foundation models function primarily as research tools, and to which extent *reliance* on these tools is an applicable concept. Subsequently, we will consider whether foundation models should be viewed as components within a broader socio-technical system, for example the organizations or companies in which they are developed, deployed and used. Trust and reliance in this case can refer to the people in these organizations, or to the organizations themselves. Finally, we will evaluate the validity of considering foundation models as research partners to which the concept of trust applies.

2. Trust versus reliance in techno-science

As indicated by John Hardwig, "those who do not trust cannot know" [20]. Against the Enlightenment ideal of the self-reliant knower which has well-founded rationales for all of her beliefs, modern knowers need to rely on the opinion of others, even in their own field of specialization. In most cases, such second-hand evidence will be even epistemically superior to knowledge based on direct evidence. It is impossible to bring all the appropriate expertise and resources (time, money, experimental power, etc.) as an individual knower to acquire knowledge in fields like particle physics or genomics. The trustworthiness of the members of the scientific community and the reliability of their research tools are therefore foundational to knowledge building. Trustworthiness is also crucial in the functioning of science in society. Members of the public need to rely on expert opinions in a huge number of diverse topics, as for instance the safety of chemical additives in food, the driving range of a car battery, or the effectiveness of a medicine for a certain ailment. This trust implies reliance on the knowledge claims as well as trust in the expertise, honesty and social responsibility of the scientists who make these claims. Trust in science therefore is an amalgamation of epistemic trust and a moral-political trust [21]. It implies having good reasons to accept a claim, as well as having good reasons to believe in the trustworthiness of the experts in the socio-technical system that gave rise to these claims.

This is increasingly true in the current techno-scientific environment, which requires a tight interplay between experts, research tools and technologies, data, the public, and governing bodies, amongst others [22]. Our practices of knowing are increasingly shaped by technologies, in particular by information technologies. Knowers in these techno-scientific environments need to continuously make decisions on which human agents are trustworthy. They also need to assess which non-human agents are reliable, for instance, companies and organisations, but also technologies such as databases, algorithms, and artificial intelligence systems. The loci of epistemic trust and reliance are, therefore, manifold and are entangled in a complex web of relations of reliance, trust and distrust [22].

Foundation models and scientific AI agents enter this epistemic fabric, as tools that can potentially be relied on, as part of a socio-technical system that potentially can be trusted or relied upon, or even as research partners that potentially can be trusted. The introduction of foundation models in research practices thereby blurs existing concepts of trust and reliance. The first reason for this is the complexity of the object of trust or reliance. From an instrumental perspective, a specific foundation model can be considered a research tool that is either reliable or not. Reliance on a foundation model then is considered in analogy with reliance on a research instrument or tool. Much like a thermometer that provides a reliable readout of the temperature, or an image analysis algorithm that allows for an accurate identification of features.

This position though is too narrow to capture what reliance and trust entail in the case of foundation models. Firstly, foundation models are the result of a multitude of decisions. One needs to decide which technologies, datasets, and training strategies can be relied upon. Scientific research instruments need to be considered within the context of the organizations in which they are developed and within the contexts in which they are used, i.e. the people in these organizations that can be trusted or distrusted.

More fundamentally, scientific foundation models capture a massive web of epistemic relations. They are trained on the results of large numbers of heterogeneous experiments executed by scientists that built their knowledge on relations of trust and reliance. When training a foundation model, an assessment needs to be made on what information sources to trust or distrust and on how to properly reflect the data from these trusted sources in the foundation model. Next to this, the scientific foundation model is trained on a body of information that incorporates a web of relations of epistemic trust and reliance. This epistemic trust or reliance is required to build hypotheses and make scientific claims. The objects of trust and reliance therefore are manifold and heterogeneous in the case of large scientific foundation models. In the case of smaller machine learning models, this object of trust and reliance is often very delineated since it considers a dedicated experimental dataset. But also machine learning models can have complex or opaque objects of trust and reliance. For example, the emergence of machine learning algorithms for facial recognition or for emotion detection highlighted the complexity of attributing responsibility [23].

Foundation models, and machine learning models in general, thus are embedded in socio-technical systems composed of many components and actors. They result from a multitude of decisions on data, technologies, training strategies, etc. And they are trained on data which is the result of relations of trust between scientists, and of relations of reliance between scientists and their instruments and data. This complexity of the object of trust and reliance can lead to blurriness of the respective concepts. This is especially the case for foundation models, given the sheer size of the data they are trained on. One can consider the foundation model as such, the socio-technical context in which the foundation model resides, the scientists that developed the model, or the scientists that provided the claims on which the model was trained. The question is how the concepts of trust and reliability apply to these different objects, which specific actions the trustee can be trusted for (or the algorithm can be relied upon), and in which context the trust or reliance applies.

The second reason for the conceptual blurriness is the very notion of reliance and trust. “Trustworthy AI” was earmarked by the European Union to frame the legal, social and technical acceptability of AI’s. Philosophically though, the concept of trustworthiness is reserved for agents that have free will and moral agency. The branding of AI as trustworthy or not therefore led to conceptual confusion. “Trustworthy AI” was considered a category mistake [24] or even conceptual nonsense [25]. The trust, in the account of Baier, “can be betrayed, or at least be let down, and not just disappointed” [26]. In this perspective, the appropriate objects of trust are the people in the institutions behind the AI, and not the AI itself. When zooming in on the AI, it is thus suggested to speak of “Reliable AI” instead [27].

The question, though, is whether one perspective on reliance or trust is sufficient to capture the complexities when foundation models enter the epistemic fabric of science. In this paper, we contrast the concept of epistemic reliance with a ‘thinner’ form of epistemic trust that builds on commitments and a ‘thicker’ form of epistemic trust that centers around shared values. This contrast allows for assessing whether mere reliance or trust is appropriate when using foundation models in science. For this analysis, we build on a pluralist account of trust in science, as put forward by Metzen [28]. This account bases itself on two main theories of trust that provide different perspectives on what exactly needs to be added to reliance to properly speak about trust.

The first theory is the goodwill account of Annette Baier. Trusting a person implies depending on the goodwill of that person to not take advantage of your vulnerability to harm “the goods or things one values or cares about” [26]. In this account, a trustworthy person is motivated by goodwill. Trust thus comprises more than reliance. It is not just the ability to rely on a certain behaviour or outcome. It includes a moral dimension, where the trustee acknowledges the values of the trustor and lets this acknowledgement guide his or her actions. Trust in this account thus implies shared values, or at least shared interests, between the trustor and the trustee.

This goodwill account underpins most theories of trust in science. It implies that one cares about the things the hearer values, next to caring about the epistemic aspects of one’s claims [28]. The theory can be applied to trust among scientists and to trust of the public in science. On the epistemic level, the focus is on what is valuable to both parties in the trust relationship. These are goods for which the hearer trusts the speaker, like knowledge, evidence and true belief, and goods for which the speaker trusts the hearer, like being recognized or acknowledged as a knower [29]. Scientific judgements though are not purely at this epistemic level. They also imply value judgements. Making scientific claims implies induction from scientific data, an activity that requires a decision on how much risk one tolerates when interpreting experimental results. A research community that shares such value judgements will make it easier for scientists to trust in the claims of others [30]. Such shared value judgements can be captured in the scientific standards of a research community. Trust in science requires taking into account the values from the hearer when making decisions on how much inductive risk to tolerate [30]. It thus implies goodwill towards fellow scientists and towards the public, in the sense of caring for their values [28].

Metzen contrasts this ‘thick’ shared values concept of trust with the ‘thin’ account proposed by Hawley [28]. This perspective on trust does not take shared values as the basis of trust, but commitments. Hawley defines trust in someone to do something, as to believe that she has a commitment to doing it [31]. Someone then is rated as untrustworthy if she cannot be relied upon to meet a commitment she made. This account of trust does not require the goodwill of the trustee. You can for example trust a pilot to safely fly the airplane you are on, without the need for her to have goodwill towards you. Trusting the pilot requires you to believe that she has a commitment to take her passengers safely to their destination, and to rely on her to live up to that commitment. This commitment is implicit in her accepting the role as a pilot. To be a trustworthy pilot can require her to be open to taking up further yet unspecific commitments, as a consequence of her initial commitment to be a pilot. These commitments can for example be the adherence to certain safety procedures, or the commitment to follow regular training. According to Metzen, such meta-commitments are key to understanding trust in science based on objectivity. Trusting a scientist in her claims is believing that the scientist adheres to the meta-commitment to produce scientific claims according to procedures that are agreed upon by her scientific community [28].

In both Baier’s account and Hawley’s account, trust differentiates from mere reliance because an extra element is added. In the commitment account, this element is the commitment of scientists to follow certain processes that are accepted by their scientific community and ensure correct epistemic claims. This perspective on trust in science is morally ‘thin’. In the shared values account, the extra elements are goodwill and the values that are shared among scientists, or among scientists and the public. This perspective on trust in science is morally ‘thick’. It for instance allows us to understand distrust in cases in which the values of individuals do not resonate with the values used in the scientific inquiry. In the pluralist perspective of Metzen, both accounts are relevant to understanding the variety of trust relations that are a play [28].

In the following sections, we will explore foundation models in science from these different perspectives. Firstly, we will analyse whether foundation models are research tools, to which the concept of reliance applies. Then, we discuss whether foundation models are to be considered elements in a large socio-technical system. Trust or reliance can then apply to the individuals in the system, or to the system itself. Lastly, we discuss whether foundation models are to be considered as research partners, that can be trustworthy or not. The applicability of the commitment account of trust and of the shared values account of trust will be explored from these perspectives.

3. Epistemic reliance and foundation models: the tool perspective

How can both accounts of trust help us to understand the impact of foundation models on the epistemic fabric of science? First and foremost, the question needs to be raised whether the relational web of epistemic trust is anyhow impacted by the introduction of foundation models. According to Jones, “Trusting is not an attitude that we can adopt toward machinery. […] One can only trust things that have wills since only things with wills can have goodwills - although having a will is to be given a generous interpretation so as to include, for example, firms and government bodies. Machinery can be relied on, but only agents, natural or artificial, can be trusted.” [32] Scientific foundation models are algorithms and thus reside on the machinery side of the spectrum. They can be components though of agentic systems that are interactive, autonomous and adapt themselves in response to novel information. Such an agentic perspective requires us to consider the relevance of *trust*, which will be discussed in Section 5 of this paper. The machinery perspective on the other hand requires an analysis of *reliance*.

Reliance is a way of acting under the assumption that the technology will perform, while trust is an attitude that implies a moral aspect. [33]. Genuine epistemic trust therefore is not at stake when interacting with a scientific foundation model that does not act like an agent, but as an algorithm that provides a response given a certain query. The question is whether the concept of epistemic reliance captures this relation. Machine learning methods, including foundation models, are commonly used in research. One way to conceptualize a scientist’s interaction with a foundation model is to view it as analogous to using a research tool, which can be relied upon to generate certain knowledge claims. Unlike relying on the testimony of another individual, reliance on a tool can be considered “morally thin.” It does not imply an agent that needs to fulfill epistemic responsibilities, such as striving to speak the truth or accurately assessing one’s own competencies.

From this perspective, epistemic reliance on foundation models could be considered akin to relying on measuring instruments (physically mediated tools) or algorithmic methods (theoretically mediated tools). Physically mediated instruments typically exploit well-understood physical phenomena that can be mathematically quantified. In contrast, theoretically mediated instruments, such as simulations or statistical analyses, involve theoretically informed procedures designed to produce reliable results [14]. Foundation models provide specific results based on the inputs they receive. They differ, though, from both physically and theoretically mediated tools. Unlike physically mediated tools, which operate on well-understood and reproducible physical phenomena, foundation models function through computational processes. For this reason, they bear similarities with theoretically mediated tools. Deep learning models are composed of massive amounts of mathematical functions and weighted connections. They have well-defined network architectures and procedural methods for executing the learning, and for mapping the inputs to the outputs. Deep learning models nevertheless differ fundamentally from theoretically mediated tools. This difference primarily stems from what is known as the epistemic opacity of deep learning models. Such models are described as epistemically opaque because it is not possible to discern all the epistemically relevant aspects of their computational processes [34]. Brute induction is not sufficient as a method to assess the reliability of their output [14]. Deep learning models are trained on specific datasets, but this does not warrant that they generalise well towards other datasets. Many deep learning models might have a similar performance on the same datasets but perform differently when used beyond this scope. The deep learning model can be very reliable, but brute induction will not put us in a position to fully assess this reliability beyond its training scope. Epistemic justification of global claims based on deep learning models is therefore complicated since one cannot trace and justify all computational steps in the model [14]. This issue is enlarged by the sheer size in the case of LLMs, and by dependencies between LLMs as they derive information from each other, for example in the process of model distillation.

A comparison of foundation models with scientific databases can shed light on the distinction between reliance on theoretically mediated tools and reliance on foundation models. Querying a database will always provide an output that can be fully traced back to individual records in the database. Deriving information from a foundation model, though, will lack this ability to trace the provenance. Consider, for example, a well-annotated and curated protein database containing extensive data on protein sequences and structures. This database serves as a research tool for scientists exploring sequence-structure relationships. A researcher can identify analogous proteins within the sequence space and deduce structural similarities. The process is transparent: for any given output, it is clear which inputs led to that result, allowing for further tracing back to the original experimental data. The database in this case is a theoretically mediated tool, the reliability of which can be assessed from knowledge of its inner workings and content. In contrast, a foundation model which is trained on exactly the same dataset would require a very different type of reliance. Unlike databases, foundation models are characterised by epistemic opacity. It is often not possible to know which inputs contributed to the generation of a specific output. For instance, a scientist might use a foundation model to predict proteins with particular structural characteristics. The relationships between the generated protein sequences and the source sequences on which the foundation model was trained are however complex and do not allow for direct traceability. XAI techniques may provide insights into which source sequences are the main contributors to the output. A direct relation though can often not be inferred.

Furthermore, foundation models are ‘epistemically promiscuous’. They have an open-ended nature regarding their potential usage, allowing for very diverse epistemic applications. This epistemic promiscuity makes that the epistemic reliability of a foundation model strongly depends on which question is asked, and how the information is extracted. One might have concluded that a foundation model functions reliably given a certain research approach, but that fact does not guarantee that it can also be relied upon when having a different scientific question and strategy for using the foundation model. Next to this effect, LLMs have the tendency to produce hallucinations. These misleading or incorrect outputs can be difficult to identify, unless a sufficient amount of ground truth data is available. These reasons make clear that epistemic reliance on foundation models differs from epistemic reliance on scientific databases, and in general, that it differs from epistemic reliance on theoretically mediated research tools.

Foundation models, as other scientific tools, require careful evaluation when used in the context of epistemic justification. Justification of claims made by the foundation models requires having sufficient reasons for asserting the validity of these claims. Such reasons need to be made according to scientific methodologies that have a proven effectiveness in producing valid claims. XAI approaches aim at elucidating the inner workings of the AI, to understand its effectiveness in producing valid claims. But next to this, justification of the validity of the claims also requires an understanding of the socio-technical perspective in which the AI is developed and deployed. More precisely, the algorithm needs to be the result of methods, algorithms, expert competencies, etc., that are formalized and proved to be effective in producing valid claims in a particular scientific field. Such an ‘externalist epistemology of algorithms’ requires that the algorithm is developed and operated in a formalized sociotechnical context, for it to be reliable in producing justified output [35]. Therefore, epistemic reliance on foundation models can only be meaningful when not only considering the algorithm, but also its sociotechnical perspective.

4. Epistemic trust and scientific foundation models: the sociotechnical perspective

When scientists employ foundation models, their epistemic engagement involves more than mere reliance on a research tool. One main reason for this is that epistemic trust in AI must ultimately refer to trust in the broader sociotechnical context that developed and deployed the AI. Such an organizational perspective on trust in AI especially makes sense when considering the practices that lead to foundation models. Foundation models most often are not trained by the scientist, nor is the training process reproducible by the scientist, given the sheer quantity of resources and the expertise that is required. The development of these models is often driven by organizations rather than by individual researchers, given this resource-intensive character. AlphaFold and GNoME for instance were both developed by Deepmind, a Google's subsidiary. Positionality effects concerning data, algorithms and computational resources, and their related data frictions, lead to systems and processes that are inherently not transparent [36]. This results in double intransparency. The reliability of the outputs of a foundation model depends on the inner workings of the model, which are partially opaque. And these inner workings are determined by how the organization that developed the model constructed and trained the model. The information on how this was done is often not or not fully disclosed by the organization, leading to another level of opaqueness. This double epistemic intransparency necessitates an epistemic relationship richer than mere reliance. It always concerns a combination of epistemic reliance on the AI, with trust in the organization and trust in the people behind the AI.

This leads us to consider another perspective on epistemic trust in AI. The object of trust might be the sociotechnical context in which the AI is embedded, rather than the AI itself. One can trust or distrust the company or organization that developed the AI, and rely (or not) on its AI products. This resonates with the viewpoint that trust in technology metaphorically refers to the humans behind the technology [37,25].

Organizations – similar to AI's - are non-human entities. This begs the question of whether trust is an appropriate concept to capture the epistemic relation between scientists and the organizations that developed the AI. In everyday life, it is common for people to have a certain level of trust or distrust in complex institutions like governments, political parties, healthcare organizations, companies, or the media. It is though the question whether trust is at stake here, or rather reliance. Hawley provided arguments for the latter point. She does not retain the notion of trust in groups, and describes our epistemic interaction with groups as a form of reliance [38]. In her account, trust is distinguished from reliance by the fact that trustees have reactive attitudes towards the trust placed in them. Trust thus implies reactive attitudes, and one can argue that groups do not have reactive attitudes in the way Strawson defined them [39]. Groups cannot meet our trust with goodwill, affection, esteem, nor with contempt, indifference, or malevolence. Only the individual group members are capable of meeting our trust in this way. In the case of foundation models, individual researchers in a large organization can have goodwill to contribute to a model that benefits academic researchers in their pursuit of knowledge or societally beneficial solutions, or they can value the fact that researcher contributed with their data to the training of the model, etc. The institutions in which these researchers function work differently. One can expect them to behave according to certain values (and to avoid certain disvalues). But they cannot be expected to hold reactive attitudes, since organizations cannot take the actions of others personally.

In response to this view, Bennet [40] proposes a commitment-based account of group trust. When we trust someone, our willingness to make ourselves vulnerable to that person's action is driven by the conviction that this person holds certain commitments that we value. Institutions can have commitments, for example, the commitment to develop foundation models that are safe or beneficial to society or that provide highly accurate scientific output. Our trust in institutions then is driven by the conviction that this group lives up to commitments that are in line with our values, and next to this, because of checks and balances. We can trust or distrust institutions because they can (or do not) act according to certain values, and live up (or don't live up) to certain commitments. Groups also can be responsive and change their course of action in case of feedback from certain parties. Which goes beyond reliance, since we do not just expect that the institution acts in a favorable way. We expect that the commitments behind these actions are in line with our values. This analysis of group trust has important consequences for how we should consider trust in foundation models. It makes clear that an organization's commitments and the values that underpin these are central to the trust in this organization and its foundation models. It is therefore important for trust to make explicit which values these are, and how well they are embedded in the processes of the organization. Such an approach allows for designing systems for trust in organizations. The increased complexity of sociotechnical systems requires putting the control at the level of organizations, rather than individual entities. This requires organizations to have their information systems supporting trust. This can be pursued via embedding values in processes, and designing processes that ensure compliance with applicable regulations and best practices [41].

Next to this, trust in an organization's digital products has a strong temporal component [42]. Trust is a relational concept, and building a relation of trust takes time. In the case of products that are used in an epistemic context, trust in an organization and the willingness to rely on its products is built over time. The initial release of AlphaFold for instance required corroboration by the scientific community to build confidence in which domains the algorithm provided reliable results, and which domains the algorithm was not accurate. Trust can be lost quickly when a major issue arises, for instance when hallucinations became apparent when testing Meta's Galaxy LLM in the context of epistemic justification [9].

The time component in trust building also highlights the importance of expertise. An expert user is someone who has built familiarity with the system, its application domains, and its limits. Non-expert users can often quickly derive results from a system, but they rely on perceptual experience rather than on hands-on experience to assess trustworthiness. As with trust in online environments, such assessment of trustworthiness then is based on, for instance, the credibility of the organisation, the ease of use of their algorithm, or the risk related to faulty outputs of the algorithm [42].

5. Scientific foundation models as trusted research partners

Each scientific enterprise is underpinned by a fabric of trust between epistemic agents. AI models increasingly enter this fabric as epistemic agents and assume roles as contributors to collective epistemic endeavours. AI agents focus on reasoning, adaptive learning, and autonomous decision-making. The agents can be made to learn, so to improve on their strategies as they encounter new situations. Foundation models are often core to the design of such agents. Such agency does not require free will or intentions, in contrast to human agency. According to some authors, agency just requires interactivity, autonomy and adaptability, defined as the ability to respond to stimuli, the ability to change states without involvement of a stimulus, and the ability to change how to respond to stimuli [43]. This agentic AI perspective raises the question of whether it is more appropriate to conceptualize this relationship as a form of trust, akin to trusting an expert [14], rather than as mere reliance on a research tool. The epistemic dependency on AI would then be characterized as a relationship of trust in an "AI epistemic partner" or an "AI research partner".

Trust in the claims made by a researcher typically hinges on the assessment of the researcher's trustworthiness. Epistemic trustworthiness implies the *ability* to provide correct information. It also implies the *will* to provide correct information, and the absence of motives to provide incorrect information. Foundation models as epistemic agents lack the latter. Epistemic trustworthiness also involves the will and the capacity to offer good reasons for the made claims [20], which means that claims are derived according to well-established research practices within their domain of expertise and that all claims ultimately are supported by first-order reasons. It is contestable though whether AI's do have such a thing as reasons for the claims they produce. It is though increasingly possible for AI agents to provide insight in the various steps in their reasoning process, for instance via CoT techniques. In this sense, AI agents can be considered to have 'good reasons' if the reasoning process is clear, if it follows procedures that are well accepted by the scientific community, and if it is underpinned with facts that can be validated. The justifications for the AI's claims are ultimately rooted in for instance the data it was trained on, the architecture of the deep learning model, and the hyperparameters used. As indicated, a rational outline of such justifications via XAI methods is feasible only on a local level, and only in certain cases [14].

For these reasons, both the 'thick' and 'thin' concepts of trust seem not to be applicable to AI agents. In Baier's account of trust, a trustworthy person is driven by goodwill, and acknowledges the values of the trustor in his or her actions. In Haley's account of trust, a trustworthy person lives up to certain commitments. AI agents cannot have genuine good will towards the trustor since they lack the capacity to have intentionality and emotions. AI agents also lack mental states, and their assertions are therefore phenomenologically different from assertions made by human experts [44]. Furthermore, at the moral level, AI agents lack free will and do not have moral intentions. They thus do not qualify as moral agents that can be held responsible for their actions. Responsibility in case of failures need to be attributed to the moral agents, which are the people that developed, deployed or applied the AI agents. Responsibility, and trustworthiness as its correlate, therefore do not have to be situated in the AI agent, but in the socio-technical system and the people that participate in it. For these reasons, it can be considered a category mistake to anthropomorphize AI's as something that has the capacity to be trusted [27,24].

The concept of reliance captures the epistemic relation with an AI agent more properly. Reliance, in contrast to trust, does not require intentionality, free will, and moral awareness. Reliance requires predictability of the outcome. I can rely on my car to stop if I use the break when parking. This reliance can have a moral aspect, though. Technological artefacts are capable of embodying values and of behaving in accordance with them [45,46]. For example, in a car equipped with sensors, I would also rely on the car to alert me if I am about to damage another car or hurt a pedestrian. Such a car is expected to embed, to a certain extent, the value of ensuring safety of both its passengers and its surroundings. Along the same lines, reliance on an epistemic AI agent can imply more than mere expectations of a certain quality of the output. This reliance can include the expectation that the epistemic AI agent embodies certain values. Instead of a 'thinner' version of trust, we therefore argue that the epistemic relation towards AI agents can be conceptualized by a 'thick' version of reliance. Reliable AI agents participate in the web of epistemic interactions as providers of information that can be relied upon, because they embed the epistemic values that scientists endorse and are designed to incorporate the procedures that are accepted in their field of science.

The central issue then revolves around the specific values and processes that are embedded in the design of the foundation model, the effectiveness with which the AI enacts on these, and how transparent the designers (can) be about those values and processes. This issue becomes increasingly significant with more interactive AIs that take part in the social epistemic fabric. They can thereby form reliance-based relationships with scientists, when considering reliance in the morally thick way as embedded values that are in line with the scientist's values. Also, in contrast to sentient human beings, one cannot ascribe commitments to technologies. The commitments are held by the designers of the technologies and the organizations in which they operate, and in the best case translated into the processes and the values embedded in the technologies. For these reasons, in the case of AI agents in science, a 'thick' concept of reliance is appropriate, rather than a thin version of genuine trust. Such a notion refers to an expectation of the predictability of the outcomes, but also embedded values that are in line with the interests and values that work with the AI agent. On the other hand, when speaking about trust in AI agents, the notion of 'trust' can only refer to an 'indirect' trust in the creators, operators or users of the AI agents. This can be 'thin' or 'thick' indirect trust along the lines of a pluralist account [28], depending on the situation.

6. Discussion

Foundation models are increasingly applied across various scientific disciplines, some even transforming entire fields of research. These models are novel research tools that require verification of their reliability in producing valid output. Consequently, the concept of epistemic reliance demands further clarification. Epistemic reliance may be justified if the mechanisms leading to knowledge claims are well-understood or are according to mechanisms that have a proven record of producing valid knowledge claims.

Moreover, foundation models introduce a novel element to the scientific discourse. Their increasingly conversational nature makes that these models are becoming an integral part of the fabric of knowledge exchange among experts. Their functioning as expert agents prompts a need to clarify the nature of trust relationships traditionally reserved for interactions among human experts. It raises the question of whether epistemic trustworthiness, rather than mere reliability, is the appropriate concept to describe these relations.

The full significance of both reliability and trust emerges only when the foundation model is considered within the sociotechnical system in which it is designed and in which it operates. The goodwill account of epistemic trust is intelligible only within this broader framework. This type of trust hinges on the capacity of the trustee to consider and align with the values of the trustor, approached with goodwill [26]. Given that artificial agents lack intentionality, this form of trust is applicable to the human creators behind the models, not to the models themselves. A thinner concept of trust might be extended to the organizations that employ these human creators. Although organizations do not possess intentionality or reactive attitudes, they can uphold commitments that are reflected in their values, procedures, mission statements, and objectives. Thus, Hawley's notion of trust as belief in a person to meeting a commitment is therefore also applicable to organizations [38]. An outline of the concepts of trust and reliance, and the entitities to which they apply in this context, is provided in Table 1.

Shifting the focus from the model as a scientific tool to the model within its sociotechnical context enables an assessment of where the notion of epistemic trust is applicable, and where reliance is a more suitable concept. Firstly, trust is a relevant concept in the interactions with the experts that develop, deploy and use the AI. Epistemic trust in this sense is the belief of the trustor that the trustee will meet her epistemic values with goodwill, or that the trustee has a commitment to meet her knowledge needs in an appropriate way. Secondly, trust is also a relevant concept in the interactions with the organization that develops the AI. In this case though, trust needs to be interpreted in the thin sense. The trustor acts in the belief that what she values is reflected in the organization's processes, goals, culture, etc., so that the organization can live up to the commitments it makes. Thirdly, epistemic trust applies to the scientific information the foundation model was trained on, since this body of information is the result of a large network of trust or distrust by scientists in the claims that other scientists make.

These three elements highlight that the openness of the scientific foundation model is an important factor in its trustworthiness. Trust in technologies implies trust in the people behind the technologies, as well

Table 1
An outline of concepts of trust and reliance, and the entities to which they apply in the field of scientific foundation models and related AI agents.

Socio-technical systems in which the scientific AI agents and foundation models are developed, deployed and used
The human agents in these systems
'Thick' account of Trust (Baier)
Trust in someone to do something, as to believe that the person is motivated by goodwill, which means that the person takes your values into account and let his or her actions be guided by this acknowledgement
'Thin' account of Trust (Hawley)
Trust in someone to do something, as to believe that the person has a commitment to doing it
Scientific AI-agents and foundation models
'Thick' account of Reliance
An expectation of the predictability of the outcomes of the algorithm, next to an expectation that values are embedded that are in line with the own interests and values, and that procedures are followed that are in line with the agreed upon procedures in the field of science
'Indirect' Trust
Trust in the integrity and the values of the creators, operators or users of the AI agents

as in the social institutions in which these technologies are developed and deployed [33]. Open-sourced models allow for more data transparency and algorithmic transparency, thereby providing better means to assess the reliability of the model. Open sourcing of the model thereby positively affects the ability to assess the trustworthiness of the organization that developed it. The willingness to open the model for scrutinization and to contribute to the public benefit can function as trustworthiness indicators.

Adopting a sociotechnical perspective also allows us to include the moral dimensions into the trust equation, next to the pure epistemic values. Moral and epistemic trust are inherently intertwined. It is therefore not always possible to trust purely at the epistemic level. For example, the interests of developers and their organizations have an impact on the degree of attention devoted to the explainability and transparency of the model, the inclusion of ethical aspects, or in the openness of the model. This perspective thus allows for the development of a 'moralized concept of epistemic trust' in the context of scientific foundation models. Such moralized concept of epistemic trust [47] includes "non-epistemic value considerations, non-epistemic norms of communication and affective trust". Such values are for instance economic justice or fair data ownership, and exemplify non-epistemic factors. Further research on the perspective of value inclusion in scientific foundation models and in the respective organizations in which they reside is required.

As AI agents increasingly and successfully participate in scientific activities, it is important to clearly distinguish where trust and reliance applies. This distinction underpins the fabric of trust between scientists, and trust between science and the broader public. As argued in this paper, trust applies to human agents and the organizations they are part of, while reliance applies to agentic AI systems. One can potentially rely on agentic AI in a 'thick' way that includes also moral perspectives. Next to the expectation of the predictability of the outcomes of the algorithm, this then also includes the expectation that values are embedded which are in line with the own interests and values, and that procedures are followed that are in line with the agreed upon procedures in the field of science. Trust, on the other hand, applies to the socio-technical systems in which the scientific AI agents and foundation models reside, and to the human agents in these systems. Such distinction between trust and reliance is necessary for the unambiguous allocation of moral responsibility to the human agents in this fabric. Maintenance and development of this fabric of trust is core to the scientific endeavour of knowledge building.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT 4o in order to review the grammar and the clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Koen Bruynseels: Writing – review & editing, Writing – original draft, Conceptualization. **Lotte Asveld:** Writing – review & editing. **Jeroen van den Hoven:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Miller B, Freiman O. Trust and distributed epistemic labor. The routledge handbook of trust and philosophy. Routledge; 2020.

[2] Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2. Aug.

[3] Lin Z, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30. https://doi.org/10.1126/science.ade2574. Mar.

[4] Krishna R, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science 2024;384:eadl2528. https://doi.org/10.1126/science.adl2528. Mar.

[5] Berman HM, et al. The protein data bank. Nucleic Acids Res 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235. Jan.

[6] Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature 2023;624:80–5. https://doi.org/10.1038/s41586-023-06735-9. Dec.

[7] Jain A, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002. https://doi.org/10.1063/1.4812323. Jul.

[8] Santos-Júnior CD, et al. Discovery of antimicrobial peptides in the global microbiome with machine learning. Cell 2024;187(14):3761–3778.e16. https://doi.org/10.1016/j.cell.2024.05.013. Jun.

[9] Chartier-Edwards N, Grenier E, Goujon V. Galactica's dis-assemblage: meta's beta and the omega of post-human science. AI & Soc 2024. https://doi.org/10.1007/s00146-024-02088-7. Oct.

[10] Jones N. AI hallucinations can't be stopped — But these techniques can limit their damage. Nature 2025;637:778–80. https://doi.org/10.1038/d41586-025-00068-5. Jan.

[11] de Winter JCF, Dodou D, Eisma YB. System 2 thinking in OpenAI's o1-preview model: near-perfect performance on a mathematics exam. Computers 2024;13(11):278. https://doi.org/10.3390/computers13110278. Nov.

[12] Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025;368:13–4. https://doi.org/10.1038/d41586-025-00229-6. Jan.

[13] Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. Ethics Inf Technol 2024;26(2):38. https://doi.org/10.1007/s10676-024-09775-5. Jun.

[14] Duede E. Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. Synthese 2022;200(6):491. https://doi.org/10.1007/s11229-022-03975-6. Nov.

[15] Callaway E. Who will make AlphaFold3 open source? Scientists race to crack AI model. Nature 2024;630:14–5. https://doi.org/10.1038/d41586-024-01555-x. May.

[16] Callaway E. AI protein-prediction tool AlphaFold3 is now more open. Nature 2024;635:531–2. https://doi.org/10.1038/d41586-024-03708-4. Nov.

[17] Schwaller F. What is science? Tech heavyweights brawl over definition. Nature 2024. https://doi.org/10.1038/d41586-024-01626-z. May.

[18] Veluwenkamp H, van den Hoven J. Design for values and conceptual engineering. Ethics Inf Technol 2023;25(1):2. https://doi.org/10.1007/s10676-022-09675-6. Jan.

[19] Hopster J, Löhr G. Conceptual Engineering and philosophy of technology: amelioration or adaptation? Philos Technol 2023;36(4):70. https://doi.org/10.1007/s13347-023-00670-3. Oct.

[20] Hardwig J. The role of trust in knowledge. J Philosoph 1991;88(12):693–708.

[21] Rolin KH. Objectivity, trust and social responsibility. Synthese 2021;199(1):513–33. https://doi.org/10.1007/s11229-020-02669-1. Dec.

[22] Simon J. Distributed epistemic responsibility in a hyperconnected era. In: Floridi L, editor. The onlife manifesto: being human in a hyperconnected era. Cham: Springer International Publishing; 2015. p. 145–59. https://doi.org/10.1007/978-3-319-04093-6_17.

[23] Crawford K. The atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press; 2021. https://doi.org/10.2307/j.ctv1ghv45t.

[24] Deroy O. The ethics of terminology: can we use Human terms to describe AI? Topoi 2023;42(3):881–9. https://doi.org/10.1007/s11245-023-09934-1. Jul.

[25] Freiman O. Making sense of the conceptual nonsense 'trustworthy AI. AI Ethics 2023;3(4):1351–60. https://doi.org/10.1007/s43681-022-00241-w. Nov.

[26] Baier A. Trust and antitrust. Ethics 1986;96(2):231–60. http://www.jstor.org/stable/2381376.

[27] Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 2020;26(5):2749–67. https://doi.org/10.1007/s11948-020-00228-y. Oct.

[28] Metzen H. Objectivity, shared values, and trust. Synthese 2024;203(2):60. https://doi.org/10.1007/s11229-024-04493-3. Feb.

[29] Dormandy K. Introduction: an overview of trust and some key epistemological applications. Trust in epistemology. Routledge; 2020. p. 1–40.

[30] Wilholt T. Epistemic Trust in science. Br J Philos Sci 2013;64(2):233–53. https://doi.org/10.1093/bjps/axs007. Jun.

[31] Hawley K. Trust, distrust and commitment. Nous 2014;48(1):1–20. https://doi.org/10.1111/nous.12000. Mar.

[32] Jones K. Trust as an affective attitude. Ethics 1996;107(1):4–25.

[33] Nickel PJ. Trust in technological systems. In: de Vries MJ, Hansson SO, Meijers AWM, editors. Norms in technology. Dordrecht: Springer Netherlands; 2013. p. 223–37. https://doi.org/10.1007/978-94-007-5243-6_14. Eds.

[34] Humphreys P. The philosophical novelty of computer simulation methods. Synthese 2009;169(3):615–26. https://doi.org/10.1007/s11229-008-9435-2. Aug.

[35] J.M. Duran, "Beyond transparency: computational reliabilism as an externalist epistemology of algorithms." Accessed: Nov. 26, 2024. [Online]. Available: https://philsci-archive.pitt.edu/23832/.

[36] Bruynseels K. Responsible innovation in synthetic biology in response to COVID-19: the role of data positionality. Ethics Inf Technol 2021;23(Suppl 1):117. https://doi.org/10.1007/s10676-020-09565-9.

[37] Pitt JC. It's not about technology. Know Techn Pol 2010;23(3):445–54. https://doi.org/10.1007/s12130-010-9125-5. Dec.

[38] Hawley K. Trustworthy groups and organizations. In: Faulkner P, Simpson T, editors. The philosophy of trust. Oxford University Press; 2017. https://doi.org/10.1093/acprof:oso/9780198732549.003.0014.

[39] Strawson P. Freedom and resentment. Proc Br Acad 1962;48:187–211.

[40] Bennett M. Trusting groups. Philos Psychol 2024;37(1):196–215. https://doi.org/10.1080/09515089.2023.2179478. Jan.

[41] Vermaas PE, Tan YH, van den Hoven J, Burgemeestre B, Hulstijn J. Designing for trust: a case of value-sensitive design. Know Techn Pol 2010;23(3):491–505. https://doi.org/10.1007/s12130-010-9130-8. Dec.

[42] Corritore CL, Kracher B, Wiedenbeck S. On-line trust: concepts, evolving themes, a model. Int J Hum Comput Stud 2003;58(6):737–58. https://doi.org/10.1016/S1071-5819(03)00041-7. Jun.

[43] Floridi L, Sanders JW. On the morality of artificial agents. Minds Mach 2004;14(3):349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d. Aug.

[44] Arora C. Proxy assertions and agency: the case of machine-assertions. Philos Technol 2024;37(1):15. https://doi.org/10.1007/s13347-024-00703-5. Jan.

[45] van den Hoven J. Value sensitive design and responsible innovation. Responsible innovation. John Wiley & Sons, Ltd; 2013. p. 75–83. https://doi.org/10.1002/9781118551424.ch4.

[46] van de Poel I. Embedding values in artificial intelligence (AI) systems. Minds Mach 2020;30(3):385–409. https://doi.org/10.1007/s11023-020-09537-4. Sep.

[47] Barimah GK. Epistemic trust in scientific experts: a moral dimension. Sci Eng Ethics 2024;30(3):21. https://doi.org/10.1007/s11948-024-00489-x. May.