Topoi
https://doi.org/10.1007/s11245-025-10164-w

How to Recognize Artificial Mathematical Intelligence in Theorem Proving

Markus Pantsar[1]

Accepted: 6 January 2025

Abstract
One key question in the philosophy of artificial intelligence (AI) concerns how we can recognize artificial systems as intelligent. To make the general question more manageable, I focus on a particular type of AI, namely one that can prove mathematical theorems. The current generation of automated theorem provers are not understood to possess intelligence, but in my thought experiment an AI provides humanly interesting proofs of theorems and communicates them in human-like manner as scientific papers. I then ask what the criteria could be for recognizing such an AI as intelligent. I propose an approach in which the relevant criteria are based on the AI's interaction within the mathematical community. Finally, I ask whether we can deny the intelligence of the AI in such a scenario based on reasons other than its (non-biological) material construction.

Keywords Artificial intelligence · Mathematics · Automated theorem proving · Mathematical proof · Philosophy of mathematical practice

1 Introduction

Among all the philosophically interesting issues concerning artificial intelligence (AI), one of the least satisfactorily treated in the modern literature is the question of *recognizing* intelligence. Ironically, this was among the first issues concerning AI that was given a systematic treatment, through Turing's notion of the "Imitation Game" (Turing 1950). What Turing proposed was an empirical test for establishing the intelligence of machines. Such "Turing tests" have a long history, but while Turing's approach was commendable in turning a somewhat ethereal question into something empirically testable, the way Turing tests have been understood may have ultimately done more harm than good. The focus of real-life Turing tests has not been to develop artificial intelligence independently and then test them in the manner proposed by Turing. Instead, researchers working in that field have focused on deceiving interrogators, creating an *appearance* of intelligence, as when competing for the Loebner prize for passing the Turing test (Warwick and Shah 2016). The unfortunate consequence of this strategy is the confused idea that the recognition of artificial intelligence is connected to deceiving the people responsible for the recognition.

In this paper, I reject that approach, while retaining the view that intelligence can be feasibly recognized through the behavior of a system. Instead of artificial tests based on imitation and deception, I want to explore what could be a non-artificial situation in which we could recognize an artificial system as intelligent. To make the problem manageable, rather than discussing the possibility of domain-general AI, my focus is on a very limited, specific domain, that of artificial *logical-mathematical* (from here on, for brevity simply "mathematical") intelligence. Ultimately, here I am interested in the following main question: if we managed to build an artificial mathematical intelligence, how could we recognize it as intelligent?

This question is of course closely connected to the question whether building such a mathematical AI is possible. While I will consider that question as well, answering it is not my principal aim. Instead, I proceed by assuming that artificial mathematical intelligence is possible, and considering what kind of system that could feasibly be. On that

✉ Markus Pantsar
markus.pantsar@gmail.com;
markus.pantsar@humtec.rwth-aachen.de

[1] Human Technology Center, RWTH Aachen University, Theaterstrasse 14, 52062 Aachen, Germany

Published online: 29 January 2025

basis, I will present as a thought experiment a scenario in which we could recognize its intelligence. In this scenario, an artificial system produces humanly interesting mathematical proofs autonomously. Moreover, it communicates these proofs in the preferred manner of human mathematicians, i.e., by writing and submitting scientific papers. If in this scenario the mathematical community accepts the artificial system as a member, I ask, would this count as an accurate recognition of mathematical intelligence? And if not, why? What could the arguments against the intelligence of the artificial system be in that scenario? Most importantly, are there any arguments that would not be a form of "meat chauvinism", i.e., limiting the domain of intelligence only to biological systems (see, e.g., Clark 2008)?

In Sect. 2, I review the relevant literature on what intelligence is and how it can be tested. Sect. 3 then connects those considerations to the question of recognizing *artificial* intelligence in particular. I move the focus specifically to mathematical intelligence in Sect. 4 and in Sect. 5, I present some current uses of AI applications in mathematics, with focus on theorem proving. Based on that, in Sect. 6, I present the scenario in which an AI can provide and communicate mathematical proofs of theorems in a human-like manner. Finally, in Sect. 7, I ask whether we could (or should) ascribe intelligence to such AI. I conclude that in such a scenario it could be difficult to motivate denying the intelligence of the AI, at least from a basis other than its material construction (namely, it not being a biological system).

Throughout this paper, I write about "artificial intelligence" and "AI applications," without assuming that the relevant applications are intelligent. In doing that, I am simply following the custom of the field. Under the concept of AI, I include the kind of things that AI researchers standardly do, such as chatbots, translation tools, image recognizing, etc. As a result of this use of terminology, I will ask questions like "is the artificial intelligence intelligent?" which may seem either trivial or misplaced. However, I hope that the reader bears with me, given that avoiding the use of standard AI terminology would make the presentation needlessly cumbersome.

2 What is Intelligence?

One of the first things that any immersion into literature on AI research reveals is that there is very little consensus on what should count as intelligence. Minsky (2006) has described "intelligence" as a *suitcase word*: instead of having one specific meaning, it is a word with several meanings that we need to unpack. The history of the notion of intelligence in psychology bears this out. The Aristotelean notion of intelligence as reason was finally turned into a quantifiable form in the 19th century when Galton measured intelligence through reaction times (Jensen 2002). This development led to the introduction of modern-type IQ tests, most prominently on the Stanford-Binet intelligence scale (Terman 1916). While IQ tests quickly gained popularity, especially in the US, psychologists could not agree what intelligence was, leading Spearman to lament that "…'intelligence' has become a mere vocal sound, a word with so many meanings that finally it has none" (Spearman 1927, p. 14).

Since then, the matter has hardly become clearer. While intelligence was treated as one phenomenon in early psychological research, with each person possessing one intelligence, Gardner (1983) famously proposed seven (or perhaps more) different types of intelligence. More recently, this idea of multiple intelligences has been largely replaced by the re-introduction of the notion of a single intelligence, although one more general than intelligence tested on the Stanford-Binet scale. This notion was originally tied closely to reasoning, problem solving and learning (Snyderman and Rothman 1987), but in modern literature, intelligence is often also tied to adaptability. The idea itself can already be found in the work of Stern (1920) and it has been proposed in many formulations (see, e.g., Gottfredson 1997). In perhaps the most famous one, Sternberg has defined intelligence as "the ability to learn from experience and to adapt to, shape, and select environments" (Sternberg 2012, p. 19).

Here I cannot go further into the psychological notion of intelligence, but let us take a closer look at the Sternberg definition from the perspective of non-human intelligence.[1] It is clear that non-human animals exhibit a great variety of abilities to learn from experience, as well as adapting to, shaping, and selecting environments (see, e.g., de Waal 2017). But could there also be artificial intelligence that fulfils Sternberg's criteria? If we accept that machines can learn (and experience), the first criterion of learning from experience is clearly fulfilled by a wide range of machine learning systems. But the question of adapting to, shaping, and selecting environments is much trickier. This kind of intelligence seems to be limited to self-propelled robots and not applicable more widely. Most importantly, computer programs that are run on stationary hardware could not be intelligent, thus rendering the vast majority of current AI systems fundamentally unfeasible as applications of genuine intelligence.

AI researchers have not been particularly helpful in providing a neat and tidy definition. In the 2016 Stanford University report on AI research, artificial intelligence was defined as "a branch of computer science that studies the properties

[1] For an overview of the history of intelligence in psychology, see (Cianciolo and Sternberg 2004).

of intelligence by synthesizing intelligence"[2]. In the 2021 updated report, an alternative definition is proposed: "artificial intelligence is about getting a machine to carry out behaviors that we think of as requiring intelligence"[3]. Given that no specification of intelligence is given, the circularity of these statements makes them unfit for definition. However, this is not seen as a weakness in the reports. In the 2016 report (p. 12), it is also said that "The lack of a precise, universally accepted definition of AI probably has helped the field to grow, blossom, and advance at an ever-accelerating pace." The underlying message is clear: if we want to define intelligence in an exact manner, we are only likely to hinder developments in AI research. This attitude could be connected to a commonly noted phenomenon related to AI research: whenever computers show some ability thought to require intelligence, rather than ascribe intelligence to the computer, the scientific community tends to conclude that the ability did not require intelligence after all. This has been the case with image recognition, playing games like Go and chess, translation, natural language processing, and all the other recent AI success stories (Mitchell 2019).

This shifting of the goalposts may suggest that it is hopeless to ever establish genuine artificial intelligence. If we one by one remove abilities from the list that comprises intelligence as they are achieved by artificial systems, we run the risk of implicitly defining intelligence as something that only humans (and to lesser degree other animals) can possess. This, however, would be a highly disingenuous position and a bad case of "meat chauvinism." Intelligence, as I understand it, has to be tied to *abilities* of systems, not to the materials that the systems are made out of, or their physical configurations. On this approach, there can be no *prima facie* reason why a silicon-based system (or other non-biological system) could not be intelligent just like carbon-based ones can be.[4] Therefore, any approach that *de facto* ends up coupling the notion of intelligence to humans as a biological species is unsatisfactory. This is the problem with many of the most famous skeptical accounts of AI, like that of Dreyfus (1992). Ultimately, they simply seem to assume human exceptionalism, perhaps extended to some non-human animals.[5]

It could be expected that expanding the scope of intelligence to include non-human animals does not improve things much. After all, animal intelligence can be used for the purposes of similar biological exceptionalism. While this may be the case, there is much to learn about the discourse on intelligence from the history of research on animal intelligence. Indeed, I contend that it offers us important insights that can be used when assessing the possibility of artificial intelligence. This is because for a long time, comparative psychology and the research on animal intelligence had to face similar obstacles as AI research does now. Any observed potentially intelligent behavior was quickly shot down and explained to be unintelligent, based on "instincts" (de Waal 2017). In recent times, the situation has radically transformed, and a wide range of animals are generally accepted to possess genuine intelligence. Importantly, this change has taken place without there being a consensus on what counts as intelligence. Instead, the underlying ethos seems to have been that we will recognize intelligence in animal behavior once we see it. I want to analyze in this paper whether this ethos could also transfer to the behavior of artificial systems, and if so, what the consequences could be.

To make the approach as fruitful as possible, we should be prepared for the scenario that artificial intelligence, like animal intelligence, may not resemble human intelligence in all aspects, and may therefore resist measurement by human standards. In particular, the connection to adaptability may need to be reassessed. While some forms of adaptability seem feasible as a criterion of intelligence, for artificial intelligence we may need to drop others, in particular those relating to shaping and selecting environments. A machine may be built for minimal such adaptability, but it could still display intelligence.

Thus, to replace the old anthropocentric stance toward intelligence, in which intelligence forms a continuum with the human at one extreme end, we need to be open to more complex intelligence spaces. This is in line with what Sloman (1984) wrote about the existence of minds. He argued that researchers should stop classifying things into those that have the essence of a mind, and those that do not. Instead, we should approach minds in a theoretical space that allows for different kinds of essence, in which the human mind is not used as the "measure of all minds" (Hernandez-Orallo 2017). The question of intelligence is closely related to the question of minds, and the structure

[2] Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence (AI100), 2016 Study Panel Report, p. 13, https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/ai100report10032016fnl_singles.pdf.

[3] *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100).2021 Study Panel Report*, p. 78 https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/AI100Report_MT_10.pdf.

[4] There have been, however, arguments to that effect, most recently by Landgrebe and Smith (2022). In this paper I will not focus on them.

[5] Of course it is possible to treat intelligence exclusively as a biological phenomenon. But on such an approach, there is not even a theoretical possibility of artificial intelligence, which makes it uninteresting for present purposes.

of the space of possible minds, as envisioned by Sloman, is likely to have a close resemblance to the structure of the space of possible *intelligences*.

What could the space of non-humanlike intelligence be like? It is impossible to envision all the kinds of processes that can be intelligent, whether biological or artificial. This makes the question schematic and theoretical. But we have no hope of making progress in answering the question if we do not have a clearer idea what could feasibly count as *recognizing* intelligence. In this, we should not be restricted by anthropocentric attitudes. However, that kind of restriction has been endemic to the study of intelligence both in comparative psychology and artificial intelligence research.

The study of animal intelligence has a curious history with numerous cases of experimental set-ups preventing the detection of intelligence (i.e., *false negatives*), but also of experiments interpreted as showing more intelligent behavior than justified (i.e., *false positives*). One instructive story of false negatives involves the study of gibbons, for a long time considered apes of lesser intelligence due to failing tool use tasks. The standard task involved using a stick lying on the floor to bring food within reach. However, it was found out that the hand of the gibbon, being an almost exclusively tree-dwelling animal, is not suited for picking up objects from the floor, as the task required. When the experiment was changed so that tools (in this case strings) were on the shoulder level, gibbons showed similar intelligence to other apes (Beck 1967; de Waal 2017). One must wonder how many experiments fail because of such "gibbon hands"; i.e., they fail to detect animal intelligence because of flawed set-ups that prevent the animal from using its characteristic abilities.

The reverse problem, that of false *positives*, is common in the study of numerical abilities in non-human animals (as well as in human infants). Arithmetical abilities have been reported in many animals, including newborn chicks (Rugani et al. 2009) and honeybees (Warren 2018). It is not possible to give this topic a detailed treatment here, but all these are cases of attributing more sophisticated cognitive abilities to the animals than needed to explain their behavior, which can be explained entirely based on the existence of the evolutionarily developed *proto-arithmetical* abilities of subitizing and estimating (see Pantsar 2024a for a detailed account).

A research guideline known as *Morgan's canon* tells us not to interpret non-human animal behavior in terms of complex cognitive abilities if it can be explained in terms of simpler ones (Epstein 1984). In the study of animal cognition, Morgan's canon sometimes seems to be forgotten, resulting in cases of false positives. However, traditionally the more important problem has been underestimating animal intelligence based on anthropocentric experimental set-ups, resulting in cases of false negatives. When moving the focus from animal to artificial intelligence, lessons need to be learned from both mistakes. We should not expect artificial intelligence to be human-like, and thus not be restricted to set-ups where we can only detect intelligence as we associate it with humans. Hence, we should always be concerned about possible "gibbon hands" that prevent us from detecting intelligence when it is present. But we should also be careful not to ascribe intelligence when other interpretations can explain the same phenomena, thus including an AI version of Morgan's canon in our theoretical considerations.

3 Recognizing Artificial Intelligence

Based on the considerations in the previous section, we need to be careful not to connect intelligence as a general phenomenon too closely to human intelligence. However, this may not be easy to achieve in practice. Both in AI development and applications, it is common to discuss intelligence in the particular sense of human intelligent activity. Indeed, this may be desirable: given the problems defining intelligence mentioned above, the best understanding of intelligence we have in particular tasks may come from our experiences as humans performing those tasks– and observing them to be performed. Yet we should not have those experiences limiting our understanding of intelligence. To avoid this problem, I propose dividing the topic of *recognizing* artificial intelligence into two parts. The first part concerns recognizing *human-like* artificial intelligence. The second part concerns recognizing artificial intelligence that may not have human-like characteristics.

Much of the literature on recognizing artificial intelligence is focused on the first part. The standard platform to start discussing the topic is the Turing test. Turing (1950) proposed an "Imitation Game," in which an interrogator aims, by asking questions, to recognize which one of two players is human and which one a computer. The amount of literature on Turing tests is vast and it remains an active topic in AI research. Here I cannot give that topic a wider treatment (see Gonçalves 2024 for a recent account), but it is important to notice that the Turing test is clearly limited to human-like intelligence. The very task given to computers that aim to pass the Turing test is to *imitate* human intelligence, which can result in programming computers with a particular type of test in mind.[6]

[6] While the Turing test is often framed in terms of deceiving the interrogator, Turing himself ruled out designing machines with the express purpose of passing the Turing test (Gonçalves 2023; Turing 1951). For problems about deception and recognizing intelligence, see (Pantsar 2025).

This focus on human-like intelligence is also a problem with other types of tests. Detterman (2011), for example, has suggested the development of IQ tests for computers. Human IQ tests will not do because, as shown by Sanghi and Dowe (2003), several standard tests can be passed by relatively simple programs. Thus, testing artificial intelligence requires other types of tests. Dowe and Hérnandez-Orallo (2012) have discussed ways to develop IQ tests for computers that are based on first principles, using algorithmic information theory. This indeed sounds like a more promising approach. In particular, on this approach, intelligence was defined as the ability to *comprehend*, which refers to identifying predominant patterns in evidence (Hernandez-Orallo 2000). This so-called *C-test* was based on finding a continuation to a sequence of letters, much like in standard IQ tests. However, unlike those tests, the C-test was designed on a purely computational basis employing formal notions of algorithmic information theory (such as Kolmogorov complexity), not a notion of intelligence in which difficulty is derived from human populations (Hernandez-Orallo 2000). The resulting test should therefore be free of human biases, while also working as a test for measuring human intelligence (Hernandez-Orallo 2017, 197–199).[7]

While the test proposed by Hernandez-Orallo is more general than traditional IQ tests, it is still clearly influenced by testing human-like intelligence. This prompts the question: could we have intelligence tests that are truly general, applicable to any type of intelligence? Given the difficulty of defining intelligence, this seems unrealistic. Consequently, the general problem of recognizing unhuman-like artificial intelligence may not be possible to tackle. This can be particularly problematic in the context of this paper, where the focus is on mathematical intelligence. Why would artificial mathematical intelligence necessarily be human-like? Indeed, some of the most famous applications of computers in proving theorems in mathematical practice, from the four-color theorem (Appel and Haken 1976) to the Kepler conjecture (Hales et al. 2017), have been distinctively unhuman-like. They have applied brute-force methods to exhaust an extremely large (but obviously finite) set of cases one by one, in a way that would be impossible for human mathematicians. Few would currently claim that such brute-force proofs show mathematical intelligence on the part of the computer, but that may be due to our anthropocentric understanding of intelligence. And even if we accept that the brute-force proofs do not display intelligence, it is feasible that an intelligent theorem proving program could be unhuman-like in its processing and/or behavior, but in some different way. In fact, this is something we should expect, given the great advantage that an AI can have over humans in its computational power. If we construct mathematical AI systems, we would certainly like to take advantage of their greater computational potential. In such cases, the intelligence of a system could be measured through, for example, its speed in processing tasks or capacity for greater computational complexity (for more, see Pantsar 2019, 2021).

However, I am skeptical about the possibility of such mathematical AI systems being widely recognized as intelligent, unless they can display mathematical ability that we associate with humans. I predict that if a mathematical AI system is accepted as intelligent, it will be due to it displaying behavior that can be associated with human-like intelligence. This does not mean that the AI behavior should be *exactly* human-like in the imitation sense of the Turing test. But it is likely that some level of resemblance of intelligent human behavior is required of the AI system for human observers to recognize it as intelligent. Hence, for the rest of the paper I will focus on the question of what kind of AI behavior could be feasibly recognized as intelligent in the field of mathematics. I will construct a scenario that gives the AI system the maximal potential for being recognized as intelligent by humans. But in order to do that, we must first understand better what mathematical intelligence in particular is understood to be.

4 What is Mathematical Intelligence?

While defining intelligence in general has proven to be a difficult topic, perhaps a definition of *mathematical* intelligence could be more straight-forward. It is clear that mathematical intelligence is traditionally seen as one important form of intelligence. In Gardner's (1983) theory of multiple intelligences, logical-mathematical intelligence was identified as one of the types of intelligence. In the definition offered by the *American Psychological Association*, logical-mathematical intelligence is characterized as follows:

> In the multiple-intelligences theory, the set of skills used in reasoning, abstraction, and numerical analysis and computation. These abilities are alleged to be relatively independent of the abilities involved in other types of intelligences.[8]

I cannot enter the discussion concerning to what degree logical-mathematical intelligence is independent of other types of intelligences, but the set of skills associated with mathematical intelligence are interesting from an AI perspective.

[7] As it turns out, however, the program of Sanghi and Dowe (2003) also did well in the C-test, thus rendering it as unfit as a single measure of computer intelligence as the traditional IQ tests (Hernandez-Orallo 2017, p. 198).

[8] https://dictionary.apa.org/logical-mathematical-intelligence.

It is not clear what is meant by "numerical analysis" in the above quotation, although clearly the meaning is different than in mathematics, where the term refers to numerical (approximate) solutions to problems in analysis. Most likely, the APA definition refers generally to analytic skills in the treatment of number systems (e.g., natural numbers, real numbers).

Could an AI possess mathematical intelligence thus characterized? At first glance, all the mentioned areas of skills would already appear to be within the range of abilities possessed by artificial systems. Automated theorem proving software, such as *E* and *Vampire*, and interactive theorem provers, such as *Mizar* and *Lean*, can conduct deductive proofs, which in humans we would consider a high and sophisticated level of reasoning. Through pattern recognition, machine learning AI systems can detect abstract qualities such as shapes and quantities (Stoianov and Zorzi 2012; Testolin et al. 2020; Pantsar 2023). Standardly used computer software like *Mathematica* and *MATLAB* can carry out a great variety of tasks concerning number systems. And, obviously, computers can compute. Thus, all four identified areas of skills fall within the scope of modern AI applications. Nevertheless, I presume that few would be ready to say that this generation of AI systems shows mathematical intelligence, any more than a pocket calculator shows intelligence. The AI applications show capacity in areas related to mathematical intelligence, but they are not (yet) considered to be intelligent.[9]

Therefore, the key word in the APA definition appears to be "skills". Whatever the AI systems are currently doing, they are not showing skills that we generally associate with intelligence. One way to formulate this argument is through the AI not *understanding* the mathematics that it is processing. This is the thinking behind two of the most famous classic anti-AI arguments, those by Searle (1980) and Dreyfus (1992). According to this line of argumentation, a computer is simply a mechanical symbol-manipulating machine which cannot be expected to understand what it is doing. Therefore, even in cases where a computer seemingly shows intelligent behavior, the argument goes, it is still as "dumb" as computers have always been. Searle makes the conceptual distinction between "weak" and "strong" AI for this purpose. The argument, in a nutshell, is as follows. Somehow the biological material of the human brain is conducive to understanding, which is central to human cognitive behavior. An AI system may be able to reproduce the behavior, but as long as it does not understand what it is doing, it only counts as weak AI. In strong AI, the system must understand what it does.

It is not possible to treat this topic here in detail (see (Cole 2020) for a treatment of Searle's Chinese room argument and its critics), but when it comes to strong AI, Searle's criticism clearly goes against the possibility of recognizing intelligence through mere observation of behavior. However, Searle's argument is also potential grounds for meat chauvinism. If understanding is associated too strongly with biological systems, the burden of proof on artificial systems may become unrealistically heavy: given that no behavioral observations are sufficient, what could be the kind of evidence that convinces the meat chauvinist of the presence of understanding in the processing of an AI system?

Nevertheless, to deal with Searle-type criticism, I am ready to make two further specifications to the present project. First, to follow Searle's distinction, I am ready to accept that the relevant notion of intelligence in the present context is in fact *weak AI*. If an AI system can reproduce human mathematical behavior on the highest levels, it would already be a remarkable achievement. Second, in case that the focus on weak AI is sufficient, I am ready to accept that this paper is about a particular *behaviorist* understanding, according to which intelligence is a matter of the behavior displayed by a system. This, of course, is the basis of recognizing intelligence in the Turing test. In the literature, assessing artificial intelligence through observing human-like behavior is often called the *Turing test approach* (see, e.g., Norvig and Russell 2021; Sect. 1.1). Given that the scenario here, for reasons explained in Sect. 3, is human-like mathematical AI, the argumentation in this paper can be understood as part of the Turing test approach. However, this choice is made for the purpose of evaluating the most feasible scenario in which a mathematical AI system could be accepted as intelligent. No further philosophical importance should be associated with the choice.

It is important to note that the behaviorist Turing test approach comes with the strength of focusing on the question of mathematical intelligence as such, and not its origins. Even though in humans there appears to be a great deal of path-dependency in the development of mathematical intelligence (see, e.g., Pantsar 2024a), we should not assume that the same holds for AI systems. Without a solid argument in place to suggest otherwise, we should not a priori deny the possibility of an artificial intelligence acquiring human-like (or perhaps some other type) of mathematical intelligence through an essentially different process. This openness to other routes to intelligent behavior is inherent to the Turing test approach, and it is crucial for the present project.

[9] This is at least the attitude of mathematicians from personal experience. To the best of my knowledge, no systematic studies on this topic exist.

5 Mathematics and AI

Mathematical intelligence may not be simply one thing and it could be exhibited in different ways. Therefore, the first question to ask regarding the recognition of mathematical intelligence is what kind of activity we could even *potentially* deem to be intelligent in artificial systems. Certainly, this would not be calculation, which seems to be paradigmatic of the kind of mechanical process that computers do without understanding what they are doing. Similarly, the functioning of the current generation of automatic theorem provers (ATP) like *E* and *Vampire*, as well as interactive theorem provers (ITP) like *Lean* and *Mizar*, is not likely be considered intelligent. What they do in most applications is take a problem as the input, consisting of a set of first-order axioms and a conjecture (a first-order formula). Then, typically using first-order logic with equality, the ATP checks whether the conjecture follows from the axioms (Voronkov 2003, p. 1607). Instead of a mere confirmation or disconfirmation of the conjecture, it is desirable that the ATP also produces a proof (in case of confirmation). However, regardless of whether a proof is produced, such software does not seem to be doing anything more intelligent than what a pocket calculator does. ATP and ITP software are often called *proof assistants* for a good reason: they are used as tools in theorem proving, but they do not provide proofs autonomously (Pantsar 2024b).

In principle, we could of course use an ATP to prove new theorems simply by having it list the propositions that follow from the axioms. However, even if we limit the output to a finite number of propositions (as we obviously must), we need to deal potentially with vast amounts of theorems and proofs. But most theorems and proofs in axiomatic systems are not interesting to human mathematicians. Hence, the real challenge for autonomous automated theorem provers is whether they can discriminate between interesting and trivial theorems and proofs. There have been some advances in discriminating between proofs based on their length (Fitelson and Wos 2001; Kinyon 2019; Veroff 2001). Progress has also been made in "proof planning", a process of replacing blind search in automated theorem proving by tactics based on (human-oriented) abstract proof plans (see, e.g., Melis et al. 2008). But for the most part ATPs have been of little use in determining what kind of proofs and theorems might be interesting to human mathematicians (for more, see Pantsar 2024b). Proof assistant software have become increasingly important for mathematical practice, and many hold optimist views concerning new software based on higher-order logics (Barendregt and Wiedijk 2005; Bentkamp et al. 2023). Ultimately, however, there seems to be little in the present generation of automated and interactive theorem provers that could feasibly be called intelligence.

But could that change if autonomous ATPs with more human-like capacities were developed? Could proving humanly interesting theorems (and specifying them as interesting) be the kind of ATP activity that is feasibly considered intelligent? Here I assume that it could, at least by parts of the mathematical community, and hence for the rest of this paper I focus on the scenario that an ATP could autonomously prove theorems that human mathematicians find interesting. Admittedly, this is a very limited view toward mathematical activity. For example, the discovery and exploration of new mathematical theories– and their connection to other theories– is undoubtedly a key aspect related to mathematical intelligence. Mathematical communication and education, in their many forms, are others. And of course already the focus only on logical-mathematical intelligence is limiting, given that reasoning and knowledge representation in AI systems is a much wider phenomenon. Nevertheless, such limitations notwithstanding, I believe that the scenario where an ATP autonomously proves interesting theorems can help us gain insight on recognizing intelligence (if any) in artificial systems. Indeed, I hope that the limitations are strengths of the scenario: instead of dealing with wider and more complex phenomena, we can focus on a more constrained setting.

In the next section, I will develop the scenario in detail, but in what follows, I mostly simply assume that the scenario is realistic. If the reader disagrees with that, I invite them to treat the scenario as a thought experiment, just like the Turing test has recently been interpreted as one (Gonçalves 2023). Ultimately, my purpose is to make a philosophical point about intelligence ascriptions, not to argue for the feasibility of a particular AI application. However, there are recent developments that give cause for careful optimism. While most current automated and interactive theorem proving software are entirely rule-based AI, and thus limited in their potential in discriminating between interesting and trivial theorems, recent progress with *neural theorem provers* (NTP) suggests new potential in automated theorem proving. Unlike a rule-based ATP, an NTP can be trained with human-created proofs to teach it tactics for completing proof steps (for an introduction, see Jenson 2023). This type of approach has been used recently with promising success, including building a library of "skills" to augment the theorem proving capacity of large language models (Lample et al. 2022; Wang et al. 2023). In addition, similar machine learning applications have shown success in related tasks, such as premise selection, which refers to the problem of finding mathematical statements that are relevant for proving a particular conjecture (Wang et al. 2017).

However, perhaps the greatest reason for optimism came in 2024 when several AI developers released reports of high performances by AI systems in solving problems of the

International Mathematical Olympiad. These include DeepMind's *AlphaGeometry* (DeepMind 2024b) and *AlphaProof* (DeepMind 2024a), OpenAI's *o1* (OpenAI 2024), and Harmonic's *Aristotle* (Weinberg 2024). In the DeepMind applications, as well as in Aristotle, a hybrid neuro-symbolic architecture is used. First a large language model is pre-trained on mathematical problems, which is used to generate possible proof steps in solving the problem. These are then processed in the rule-based system Lean and successful steps are used to reinforce the model (DeepMind 2024a).

So far the systems have found success in solving mathematical problems in competitions (reaching the silver-medal level in the Olympiad), but it is feasible that similar hybrid architecture could also be used in research mathematics, including theorem proving. Such automated theorem provers could learn mathematics from datasets of humanly interesting proofs and theorems, thereby recognizing patterns that could correspond to the human categorization of theorems and proofs into interesting and trivial (Pantsar 2024b). This classification between humanly interesting and trivial is difficult to explicate, so suggesting any strict criteria would be unfruitful. Aside from some simple rules, such as theorems of the form "*A* if and only if *A"* being trivial, formal rules are of limited use.

Indeed, what counts as interesting mathematics has become an important topic in what is called philosophy of mathematical practice (Mancosu 2008). For proofs, "insightfulness" has been suggested as a criterion (Macbeth 2012; Weber 2010). Another suggested criterion has been "beauty" (Johnson and Steinerberger 2019; Rota 1997), a concept that has been established to have cross-cultural similarity (Sa et al. 2024). Thomas (2017) has argued that being interesting should itself count as an aesthetic value in mathematics. To make matters even more difficult, even the notion of mathematical rigor in practice is not a completely formal one (Antonutti Marfori 2010; Avigad 2020). It is not possible to go here into the details of these research directions, but the range of problems already makes it clear that finding formal rules for what makes mathematics interesting for humans is an extremely difficult pursuit. However, with machine learning systems implicit patterns could be established in the training data. In practice, creating sufficient training data may prove to be difficult (see Pantsar 2024b). Here I only assume that some feasible set of criteria could be (implicitly) present in automated theorem provers trained on humanly constructed proofs.

6 How to Recognize Artificial Mathematical Intelligence: The Scenario

In the previous section, I laid out a scenario in which an AI application based on machine learning (perhaps in combination with a rule-based system) can provide proofs of new, humanly interesting theorems. Estimating the intelligence of such an application could lead to false positives. Theorem proving as such is not seen as an action necessarily requiring intelligence– unlike in the early days of AI, when the program *Logic Theorist* (Newell et al. 1957), for example, was seen as intelligent (Simon 1991)– but perhaps an AI that can distinguish between interesting and uninteresting theorems and proofs would be different. However, just because the AI could do that does not mean that it would be accepted as intelligent. To see why, we can consider the current AI applications for translations and text generation, such as *DeepL*, *Google Translate* and *ChatGPT*. Few philosophers are ready to call these types of AI applications intelligent. Their functioning is based on pre-trained large language models, which is made possible by the statistical detection of patterns. Instead of being intelligent, they are often described as "stochastic parrots" that function by "haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning" (Bender et al. 2021, p. 617).

If the model behind a theorem proving AI would be comparable to the current generation of large language (and multimodal) models, it would indeed seem that any intelligence ascribed to it would be a false positive. However, some AI researchers believe that pre-trained language models show potential for developing more general forms of AI which could model sensory experience also outside language (Manning 2022). Among some AI developers, there is optimism that a more general AI is within our reach soon (see, e.g., Altman 2023). There are reasons to be skeptical of the possibility of truly domain-general artificial intelligence, at least in the near future, but at the same time we should not be limited to considerations of AI based on the current generation of large language models and their applications. Indeed, already the hybrid neuro-symbolic AI systems, like the one behind AlphaProof, are different from the "stochastic parrots". In the rule-based part of the system, the generated suggestions are put under a strict logical test, which can enable avoiding the kind of errors in reasoning tasks that large language models are notorious for (Marcus 2024). Granted, this does not imply that the systems function based on meaning, but it is conceivable that also in this respect future AI applications are different.[10]

[10] For more on the state-of-the-art, see (Bowman 2023; Zhao et al. 2023). For a skeptical view, see (Landgrebe and Smith 2022).

Above, I have remarked that theorem proving AI applications could cause false positives in ascribing intelligence. However, could it also cause false negatives, i.e., could there be a plausible scenario in which the AI application actually shows intelligence, but we are not able to recognize it? What could such a scenario plausibly be, i.e., what could a genuinely intelligent theorem proving AI be like? To give the AI system the best possible chance of being recognized as intelligent– i.e., to avoid false negatives– I believe that its artificial nature should be hidden from the evaluators. However, the scenario I am developing is not a variant of the Turing test. Instead of artificial situations where human interlocutors actively try to reveal the artificial nature of a system, we should consider scenarios in which the artificial systems are embedded into the practice of the particular field of intelligent activity. The reason for rejecting the Turing test setting is two-fold. First, it is not designed to work for domain-specific intelligence. For mathematics questions, the AI could appear intelligent even if it were not, and for non-mathematical questions it could be easily detected as artificial. Second, the Turing test method– venerable as it is in some circles– seems to be rather inconsequential in the wider perspective of artificial intelligence research. Computer software have in fact already been reported as having passed Turing test set-ups and rather than accepting them as intelligent, experts have concluded that the test in fact failed.[11]

Due to these weaknesses of the Turing test set-up, I propose a more "organic" approach for recognizing artificial intelligence. In the case of mathematical proofs, I propose that the best approach is to develop an AI system that can develop proofs that are essentially human-like. As argued in Sect. 3, if we are likely to recognize an AI system as intelligent, this will most probably happen when its behavior is human-like. This means that the AI would prove humanly interesting theorems and communicate them in a human-like way.[12] Under this conception, such a theorem proving AI would then need to have at least four characteristics. First, the theorem prover would obviously need to correctly prove theorems of some area of mathematics. Second, those theorems would need to be humanly interesting, i.e., similar enough in content to the kind of theorems human mathematicians prove. Third, the AI would need to be able to present the formal proofs in a human-like fashion. This means that instead of a line-by-line derivation of the theorem, the generated proof would consist of the key deductive steps. And fourth, the AI would need to communicate the proof in a human-like fashion, i.e., write a paper according to the style and standards of human mathematicians.

The first characteristic should be realistic to develop, perhaps most likely in a hybrid system that applies also rule-based AI. The second characteristic is potentially a more difficult prospect. The AI model would need to be trained with humanly interesting theorems and proofs, but there may not be enough of them to make the training feasible. However, it is plausible that researchers can create ways to train AI models with smaller datasets, as well as create ways of generating theorems that are humanly interesting up to a sufficient level to help with the training of the model (for details, see Pantsar 2024b).

The third characteristic is again a challenge for machine learning systems, partly for the same reasons as the second characteristic. There may not be enough existing proofs for a machine learning system to detect strong enough patterns in the ways human formal proofs are presented. However, it is plausible that progress could be made also in this regard, perhaps by generating new "human-like" proof datasets from variations of existing proofs.[13] The fourth characteristic, which may have even seemed to be the most difficult one just a couple of decades ago, seems to be quite realistic to attain. Natural language generation has improved dramatically in the past years, to the point that it is perfectly feasible that an AI application can turn a formal mathematical proof into a submittable mathematical paper.

7 How to Recognize Artificial Mathematical Intelligence: The Assessment

Now that we have detailed the scenario of an AI-generated mathematical paper that provides a proof of a new humanly interesting theorem, we can turn to the question of assessing the AI in terms of intelligence. If an AI system would submit such a paper to an esteemed mathematical journal, and get it accepted, should the system be deemed intelligent? I think that this naïve version of the assessment scenario is lacking. Perhaps an AI-generated paper could be accepted, but if it is a one-off, the reason for the acceptance might not be the intelligence of the AI. The review process, for example, may have malfunctioned. However, if an AI application would regularly produce such papers and they were

[11] "Google's AI passed a famous test — and showed how the test is broken", *Washington Post*. https://www.washingtonpost.com/technology/2022/06/17/google-ai-lamda-turing-test/.

[12] For maximal progress in mathematics, we may not want the AI to prove theorems that resemble human mathematics. In the present scenario, however, this would come with the potential difficulty of humans not being able to recognize the theorems as progress in mathematics.

[13] In this context, however, it is important to ask how we come to accept mathematical proofs in general. Viteri and DeDeo (2022) have provided an analysis of this phenomenon through "epistemic phase transitions," stages in the epistemic process of accepting an inference as valid. This kind of approach could be fruitful also for explaining how we come to accept AI-generated proofs.

accepted for publication in mathematical journals, could it actually be deemed intelligent? Or, indeed, *should* it?

In this paper, I do not want to propose a definite answer to that question. Rather, I am arguing for a kind of *community-based* approach to recognizing artificial intelligence as the way to make the scenario as fair as possible for the AI system. If the mathematical community would accept such an artificial system as a contributor, without initially knowing that it is an AI system, I claim that it would put it in the best position for its possible intelligence to be recognized by humans. There is a Zulu saying: "a person becomes a person through other persons". In my approach, mathematical intelligence would be seen in a similar way: a mathematical intelligence becomes a mathematical intelligence through other mathematical intelligences.

What could this community-based approach be like in practice? Clearly the community should be organized so that an AI would have the possibility to be included in it. Given that the mathematical community currently consists of humans who almost always have university affiliations, some changes– like a wider adoption of double-blind peer review– to the criteria would be necessary. But assuming that the necessary changes could be made, in this scenario an AI system could be introduced to the community, and it could produce mathematical research papers indistinguishable in style and in general content from humanly produced papers (for more details, see Pantsar 2025).

The question is then whether in such a scenario we could still feasibly doubt the intelligence of the artificial system. If by intelligence we mean *general* intelligence, we clearly could. It is possible– indeed, highly probable– that such a mathematical AI would be designed and trained solely based on mathematics-specific aims. Depending on the actual application, it could be that the system would have no intelligence at all in other domains. In the way artificial intelligence is discussed, the question of "real" AI is often interpreted to refer to domain-general intelligence. However, this is hardly a fair and fruitful approach. It is not fair (for the purposes of ascribing intelligence) because the challenge of building a genuine artificial intelligence may be a domain-specific challenge. That we could reveal a mathematical AI to be completely dumb concerning, say, image recognition is asking for too high criteria.

To understand artificial intelligence necessarily as domain-general intelligence is not fruitful either, because it makes the entire distinction between intelligent and unintelligent systems unapplicable in practice, at least for the foreseeable future. Building a general artificial intelligence is still far beyond the current means. Yet I believe that even in the current state of technology, it is fruitful to discuss intelligent and unintelligent systems. The entire point behind views such that the large language models behind the present-day natural language processing systems are "stochastic parrots" is that there is something in the processing of the system that is unintelligent, even if its responses may appear intelligent. But this point would hardly be relevant if we thought that only general AI can be truly intelligent, because none of the systems currently or in the near future are likely to be domain general.

I see no reason why we should assess artificial mathematical AI systems in terms of what they are *not* designed to do. If they do what their purpose is, that should be enough for assessing their (domain-specific) intelligence. If an artificial system could take its place in the mathematical community in the way discussed in this paper, I contend that we should not a priori deny its intelligence. A more fruitful approach would be to (at least provisionally) accept its possible intelligence and then discuss different types of artificial intelligence, just like cognitive and comparative psychologists have discussed different types of human intelligence and animal intelligence. What I have argued for here is that we should make an effort to avoid the potential "gibbon hand" that shadows a great deal of current AI discussion: the assumption that all intelligence is of the type that humans have. This means being open to the possibility that mathematical intelligence, even when exhibited through human-like behavior, could be based on fundamentally different processes.

Of course, also in the field of mathematics, artificial intelligence could mean a significant departure from human standards, in which case it could be even harder for humans to recognize its intelligence. In the scenario I have presented, I have focused on humanly interesting mathematics because I wanted to propose a situation in which humans would be the most likely to recognize a mathematical AI as being intelligent. The important question then is whether in that scenario there would be any other reason beyond "meat chauvinism" to deny its intelligence, i.e., only denying intelligence because it is not a biological system.

From the perspective of the internal processing of the system, mirroring Searle's Chinese room argument, such an objection could be made. Regardless of the particular type of AI architecture, the computer is still essentially doing symbol-manipulation of the type that Searle described. Hence, if understanding is necessary for intelligence, and symbol-manipulation is not thought to be conducive to understanding, any scenario of the type that I have described in this paper can be automatically classified as unintelligent. But why could understanding not emerge from symbol-manipulation? After all, understanding *does* emerge from neuronal activity, which in itself appears to be as unintelligent as symbol-manipulation in computers. This kind of counter-argument appears to fall back to meat chauvinism, making it even in principle impossible for an artificial system to be

intelligent. While such a move can be made– i.e., we can decide to treat intelligence as a purely biological phenomenon– it does not solve the fundamental problem. Instead of intelligence, we would simply need other vocabulary to discuss the AI capacities.

However, I do not want to claim that the focus on internal processing as a criterion of intelligence is itself mistaken. In this paper, I have focused on the behaviorist approach to artificial intelligence, but that was a choice made to fit the scenario. I do believe that the internal processing of AI systems should be considered in making intelligence ascriptions. Traditional automated theorem provers, for example, function based on simple rules, which is generally not considered to be intelligent activity. But can we make such a principled claim for the kind of neuro-symbolic AI architectures that AlphaProof, for example, represents? The modern versions may be flawed and limited, but is there something in their functionality that speaks against the possibility of intelligence? Detecting patterns in training data, generating possible proof steps, testing the steps based on logical rules, reinforcing the model based on the success of that testing– the combination of all this results in activity that is very different from the traditional rule-based systems.

Could that activity be intelligent? I do not want to take a stand on that issue. Any such intelligence ascription would require a much more definite understanding of intelligence than we currently have. In addition, we should also aim to have a better understanding of how the AI system functions. But if the kind of AI system that I present in the scenario above is developed, I contend that we should be ready to assess its intelligence in an unbiased fashion, independently of what earlier computer systems used in mathematics and elsewhere have been like. As a part of that assessment, the kind of test that I have introduced in this paper can be valuable because it is designed to give the AI system a fair chance of being recognized as intelligent.

8 Conclusion

In this paper, I have presented as a thought experiment a scenario in which we could assess the intelligence of AI systems when it comes to proving mathematical theorems. In the community-based approach I have suggested, the AI application could be considered intelligent if it manages to successfully contribute to the mathematical community by submitting humanly interesting proofs through mathematical papers.[14] Or perhaps it would still not be considered intelligent, in which case my question is whether this would only be due to it not being a biological system. I hope that through this analysis, the present paper can contribute to clarifying what we mean by intelligence in domain-specific cases like mathematics, and how artificial intelligence– if any– could be recognized.

However, a couple of possible counterarguments could be presented. First, while the AI system presented in the scenario clearly would be very capable in mathematics, would it still not be just a "stochastic parrot", devoid of any reference to meaning? It could exhibit intelligent-looking behavior, just like the Logic Theorist did back in 1956, but its internal processing would also be revealed as unintelligent upon closer analysis. Indeed, this is possible. If we limit ourselves to the behaviorist approach to AI, that would not be a problem, but I accept that we should also be sensitive to the functioning of the AI systems. However, this may not be so simple as in the case of rule-based systems. As AI systems become more complex and capable, it may not be easy to determine what its internal processing is like. Already in machine learning systems we lack the ability to explain the exact causes for a particular output. We should be not fooled into thinking that it will always be a straightforward matter to separate processing that can be conducive to intelligence to processing that cannot. Consequently, I believe that behavior-based recognition will be an important factor in the future.

Another counterargument that could be presented is that my scenario is just one form of the Turing test, and as such it suffers from all the weaknesses associated with that test. While there are similarities– most obviously hiding the identity of the AI system and only assessing its behavior– there are also important differences. First, in my scenario the human participants do not know that they are being tested. In the Turing test, the human interrogator is aware that one of the players is an artificial system. In my proposed test, the mathematical community is not informed of being tested.[15] Second, due to the community setting, in my scenario the AI system is assessed based on similar criteria as human members of the community. It is not actively interrogated with the purpose of determining its identity, with potential focus on its weaknesses. Rather, it can (potentially) display its intelligence through its strengths. Third, due to the first two factors, the purpose of the AI system is not to *deceive* the human interrogators as in the Turing test. Instead, the purpose is to provide advances in mathematics, which is a similar task to that of human mathematicians.[16]

While in this paper I have focused on the question of recognizing artificial mathematical intelligence, ultimately that would hardly be the most important matter in the kind of

[14] This can be extended to also communicating about the papers– and perhaps to choose where to submit the papers in the first place. Peer-review, for example, provides a particular interesting challenge.

[15] In practice, this would come with potential ethical issues that need to be resolved.

[16] For more on the details of the proposed test, see Pantsar 2025.

scenario that I have proposed. That question is interesting for philosophers, but for mathematicians the more interesting question would likely be how to *benefit* from such AI applications. Indeed, from the perspective of working mathematicians, it could well be that the question of intelligence is of secondary importance. On the more practical side of mathematical communities, we should be prepared for the ways that human agents would be applying new AI tools. For example, the development of increasingly developed automated theorem provers opens potential for the misuse of the applications. A struggling mathematician with access to such a theorem prover might generate new proofs and claim them as their own, which could create an uneven field of competition.

On the other hand, if such tools were available, we can assume that they would become widely used. Indeed, if AI tools can help mathematics progress, it would be difficult to motivate their prevention in mathematical practice. Hence the question of developing mathematical AI is tightly connected to ethical issues concerning mathematical practice. How should AI contributions be acknowledged in human-AI hybrid work? How could we detect cases of fraud in which AI contributions are presented as the author's effort? In addition to the question of recognizing artificial mathematical intelligence, many such questions need to be considered soon, as new mathematical AI applications are introduced.

Acknowledgments This paper is based on research conducted as a Senior Fellow at the Käte Hamburger Kolleg "Cultures of Research", RWTH Aachen. I want to thank the fellows and staff there for helpful discussions. I presented an earlier version of this work at the "Infinity and Intensionality" seminar, University of Oslo. I am thankful for the comments I received there. Finally, I want to thank Regina Fabry for very helpful discussions on the topic.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interest The authors has no conflicts of interest.

References

Altman S (2023) *Planning for AGI and beyond*. https://openai.com/blog/planning-for-agi-and-beyond

Antonutti Marfori M (2010) Informal proofs and mathematical rigour. *Studia Logica: Int J Symbolic Log 96*(2):261–272

Appel K, Haken W (1976) Every planar map is four colorable. *Bull Am Math Soc 82*(5):711–712

Avigad J (2020) Reliability of mathematical inference. *Synthese 198*(8):7377–7399. https://doi.org/10.1007/s11229-019-02524-y

Barendregt H, Wiedijk F (2005) The challenge of computer mathematics. *Philosophical Trans Royal Soc A: Math Phys Eng Sci 363*(1835) Article 1835. https://doi.org/10.1098/rsta.2005.1650

Beck BB (1967) A study of Problem solving by Gibbons. *Behaviour 28*(1/2):95–109

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bentkamp A, Blanchette J, Nummelin V, Tourret S, Vukmirović P, Waldmann U (2023) Mechanical mathematicians. Commun ACM. https://doi.org/10.1145/3557998

Bowman SR (2023) *Eight Things to Know about Large Language Models* (arXiv:2304.00612). arXiv. https://doi.org/10.48550/arXiv.2304.00612

Cianciolo AT, Sternberg RJ (2004) Intelligence: a brief history. Blackwell Pub

Clark A (2008) Pressing the Flesh: a tension in the study of the embodied, embedded mind?*. *Philos Phenomenol Res 76*(1):37–59. https://doi.org/10.1111/j.1933-1592.2007.00114.x

Cole D (2020) The Chinese Room Argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/chinese-room/

de Waal FBM (2017) *Are we smart enough to know how smart animals are?*

DeepMind (2024b, October 9) *AlphaGeometry: An Olympiad-level AI system for geometry*. Google DeepMind. https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/

ZDeepMind (2024a, October 9) *AI achieves silver-medal standard solving International Mathematical Olympiad problems*. Google DeepMind. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

Detterman DK (2011) A challenge to Watson. *Intelligence 39*(2):77–78. https://doi.org/10.1016/j.intell.2011.02.006

Dowe DL, Hernández-orallo J (2012) *IQ tests are not for machines, yet*

Dreyfus HL (1992) What computers still can't do: a critique of artificial reason. MIT Press

Epstein R (1984) The principle of parsimony and some applications in psychology a principle of parsimony. *J Mind Behav*, *5*

Fitelson B, Wos L (2001) Finding missing proofs with automated reasoning. *Studia Logica: Int J Symbolic Log*, *68*(3), Article 3.

Gardner H (1983) Frames of mind: The theory of multiple intelligences. Basic Books

Gonçalves B (2023) The Turing Test is a thought experiment. *Mind Mach 33*(1):1–31. https://doi.org/10.1007/s11023-022-09616-8

Gonçalves B (2024) The turing test argument. Routledge. https://www.routledge.com/The-Turing-Test-Argument/Goncalves/p/book/9781032291574

Gottfredson LS (1997) Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence 24*(1):13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

Hales T, Adams M, Bauer G, Dang TD, Harrison J, Hoang LT, Kaliszyk C, Magron V, Mclaughlin S, Nguyen TT, Nguyen QT, Nipkow T, Obua S, Pleso J, Rute J, Solovyev A, Ta THA, Tran NT, Trieu TD,…, Zumkeller R (2017) A formal proof of the kepler conjecture. *Forum Math Pi 5*. https://doi.org/10.1017/fmp.2017.1

Hernandez-Orallo J (2000) Beyond the turing test. *J Log Lang Inform 9*(4):447–466

Hernandez-Orallo J (2017) *The Measure of All Minds: Evaluating Natural and Artificial Intelligence* (1st edition). Cambridge University Press

Jensen AR (2002) Galton's legacy to research on intelligence. *J Biosoc Sci 34*(2):145–172. https://doi.org/10.1017/s0021932002001451

Jenson D (2023, May 16) Automated theorem proving with graph neural networks. *Stanford CS224W GraphML Tutorials*. https://medium.com/stanford-cs224w/automated-theorem-proving-with-graph-neural-networks-49c091024f81

Johnson SGB, Steinerberger S (2019) Intuitions about mathematical beauty: A case study in the aesthetic experience of ideas. *Cognition 189*:242–259. https://doi.org/10.1016/j.cognition.2019.04.008

Kinyon M (2019) Proof simplification and automated theorem proving. *Philosophical Trans Royal Soc A: Math Phys Eng Sci 377*(2140) Article 2140. https://doi.org/10.1098/rsta.2018.0034

Lample G, Lachaux M-A, Lavril T, Martinet X, Hayat A, Ebner G, Rodriguez A, Lacroix T (2022) *HyperTree Proof Search for Neural Theorem Proving* (arXiv:2205.11491). arXiv. https://doi.org/10.48550/arXiv.2205.11491

Landgrebe J, Smith B (2022) *Why Machines Will Never Rule the World: Artificial Intelligence without Fear* (1st edition). Routledge

Macbeth D (2012) Proof and understanding in mathematical practice. *Philosophia Scientiæ Travaux D'histoire Et de Philosophie Des Sci 16*–1(Article 16–1). https://doi.org/10.4000/philosophiascientiae.712

Mancosu P (ed) (2008) *The Philosophy of Mathematical Practice* (1st edition). Oxford University Press

Manning C (2022), April 13 *Human Language Understanding & Reasoning*. American Academy of Arts & Sciences. https://www.amacad.org/publication/human-language-understanding-reasoning

Marcus G (2024), August 1 This one important fact about current AI explains almost everything [Substack newsletter]. *Marcus on AI*. https://garymarcus.substack.com/p/this-one-important-fact-about-current

Melis E, Meier A, Siekmann J (2008) Proof planning with multiple strategies. *Artif Intell 172*(6):656–684. https://doi.org/10.1016/j.artint.2007.11.004

Minsky M (2006) The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind. Simon & Schuster

Mitchell M (2019) Artificial Intelligence: A guide for thinking humans. Illustrated edition. Farrar, Straus and Giroux

Newell A, Shaw JC, Simon HA (1957) Empirical explorations of the logic theory machine: A case study in heuristic. *Papers Presented at the February 26–28, 1957, Western Joint Computer Conference: Techniques for Reliability*, 218–230

Norvig P, Russell S (2021) *Artificial Intelligence: A Modern Approach, Global Edition* (4th edition). Pearson

OpenAI (2024) *Introducing OpenAI o1*. https://openai.com/index/introducing-openai-o1-preview/

Pantsar M (2019) Cognitive and computational complexity: considerations from mathematical problem solving. Erkenntnis 86:961–997. https://doi.org/10.1007/s10670-019-00140-3

Pantsar M (2021) Descriptive complexity, computational tractability, and the logical and cognitive foundations of mathematics. Minds and Machines 31(1):75–98. https://doi.org/10.1007/s11023-020-09545-4

Pantsar M (2023) Developing artificial human-like mathematical intelligence (and why). Minds and Machines 33:379–396

Pantsar M (2024a) Numerical Cognition and the Epistemolofy of Arithmetic. Cambridge University Press. https://doi.org/10.1017/9781009468862

Pantsar M (2024b) Theorem proving in artificial neural networks: new frontiers in mathematical AI. European Journal for Philosophy of Science 14(1):4. https://doi.org/10.1007/s13194-024-00569-6

Pantsar M (2025) Intelligence is not deception: from the turing test to community-based ascriptions. AI & Society. https://doi.org/10.1007/s00146-024-02172-y

Rota G-C (1997) The phenomenology of Mathematical Beauty. *Synthese 111*(2):171–182

Rugani R, Fontanari L, Simoni E, Regolin L, Vallortigara G (2009) Arithmetic in newborn chicks. *Proc Royal Soc B: Biol Sci 276*(1666):2451–2460

Sa R, Alcock L, Inglis M, Tanswell FS (2024) Do mathematicians agree about Mathematical Beauty? *Rev Philos Psychol 15*(1):299–325. https://doi.org/10.1007/s13164-022-00669-3

Sanghi P, Dowe DL (2003) A computer program capable of passing I.Q. tests. In *4th International Conference on Cognitive Science (ICCS'03)* (pp. 570–575). International Association for Cognitive Science

Searle JR (1980) Minds, brains, and programs. *Behav Brain Sci 3*(3):417–424. https://doi.org/10.1017/S0140525X00005756

Simon HA (1991) Models of my life. Basic Books

Sloman A (1984) The structure of the space of possible minds. In: Torrance S (ed) The mind and the machine: philosophical aspects of Artificial Intelligence. Ellis Horwood, pp 35–42

Snyderman M, Rothman S (1987) Survey of expert opinion on intelligence and aptitude testing. *Am Psychol 42*(2):137–144

Spearman C (1927) *The abilities of man* (pp. xxiii, 415). Macmillan

Stern W (1920) *Die Intelligenz der Kinder und Jugendlichen und die Methoden ihrer Untersuchung.* Barth

Sternberg RJ (2012) Intelligence. *Dialog Clin Neurosci 14*(1):19–27

Stoianov I, Zorzi M (2012) Emergence of a'visual number sense'in hierarchical generative models. *Nat Neurosci*, *15*(2), Article 2.

Terman LM (1916) *The measurement of intelligence* (pp. xiii, 374). Houghton, Mifflin and Company. https://doi.org/10.1037/10014-000

Testolin A, Zou WY, McClelland JL (2020) Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Dev Sci*, *23*(5), Article 5

Thomas RSD (2017) Beauty is not all there is to aesthetics in mathematics†. *Philosophia Mathematica 25*(1):116–127. https://doi.org/10.1093/philmat/nkw019

Turing AM (1950) Computing machinery and intelligence. *Mind LIX*(236):433–460. https://doi.org/10.1093/mind/LIX.236.433

Turing AM (1951) Can digital computers think? In: Copeland BJ (ed) The essential turing: the ideas that gave birth to the computer age. Oxford University Press, pp 482–486. https://doi.org/10.1093/oso/9780198250791.003.0019

Veroff R (2001) Finding shortest proofs: An application of linked inference rules. *J Automated Reasoning 27*:123–139

Viteri S, DeDeo S (2022) Epistemic phase transitions in mathematical proofs. *Cognition 225*:105120. https://doi.org/10.1016/j.cognition.2022.105120

Voronkov A (2003) Automated reasoning: Past story and new trends. *IJCAI*

Wang M, Tang Y, Wang J, Deng J (2017) Premise selection for theorem proving by deep graph embedding. *arXiv:1709.09994 [Cs]*. http://arxiv.org/abs/1709.09994

Wang H, Xin H, Zheng C, Li L, Liu Z, Cao Q, Huang Y, Xiong J, Shi H, Xie E, Yin J, Li Z, Liao H, Liang X (2023) *LEGO-Prover: Neural Theorem Proving with Growing Libraries* (arXiv:2310.00656). arXiv. https://doi.org/10.48550/arXiv.2310.00656

Warren M (2018) Bees understand the concept of zero. *Science*. https://www.science.org/content/article/bees-understand-concept-zero

Warwick K, Shah H (2016) Can machines think? A report on turing test experiments at the Royal Society. *Journal Experimental Theoretical Artif Intelligence 28*(6):989–1007. https://doi.org/10.1080/0952813X.2015.1055826

Weber K (2010) Proofs that develop insight. *Learn Math*, *30*(1), Article 1.

Weinberg J (2024, July 11) *New AI Venture Focuses on Mathematical Reasoning—Daily Nous*. https://dailynous.com/2024/07/11/new-ai-venture-focuses-on-mathematical-reasoning

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