Ethical guidelines for AI use in mathematical research

Preprint

Markus Pantsar (RWTH Aachen University)

Abstract

Generative artificial intelligence (AI) applications based on large language models have not enjoyed much success in symbolic processing and reasoning tasks, thus making them of little use in mathematical research. However, recently DeepMind's AlphaProof and AlphaGeometry 2 applications have recently been reported to perform well in mathematical problem solving. These applications are hybrid systems combining large language models with rule-based systems, an approach sometimes called neuro-symbolic AI. In this paper, I present a scenario in which such systems are used in research mathematics, more precisely in theorem proving. In the most extreme case, such a system could be an autonomous automated theorem prover (AATP), with the potential of proving new humanly interesting theorems and even presenting team in research papers. The use of such AI applications would be transformative to mathematical practice and demand clear ethical guidelines. In addition to that scenario, I identify other, less radical, uses of generative AI in mathematical research. I analyse how guidelines set for ethical AI use in scientific research can be applied in the case of mathematics, arguing that while there are many similarities, there is also a need for mathematics-specific guidelines.

Keywords

Artificial intelligence; automated theorem proving, mathematical AI, AI ethics, research ethics

1. Introduction

There is denying the importance of artificial intelligence (AI) applications in the modern world. With the introduction of deep neural networks and transformer architectures, machine learning systems have been successful in many areas where computers were previously of limited use. A particularly exciting development has been that of generative AI, based on large language models (LLM) and more recently multi-modal large language models. The rise in both quantity and quality of AI-generated content includes text-based systems like chatbots and translation tools, but also increasingly other media, like images, video and audio.

However, for all their success in the aforementioned fields, generative AI applications were for a long time notoriously bad in tasks involving symbolic processing and reasoning. Famous examples of that include failing basic arithmetical operations and giving false answers to simple questions like how many letters "r" there are in the word "strawberry". One particularly telling problem has been that of simple reasoning tasks that resemble widespread, more complex, tasks. In one task, the Microsoft chatbot Copilot was asked to solve the following puzzle: "A man and his goat are trying to cross a river. They have a boat. How do they do it?" Copilot first gave a five-part solution including taking a cabbage across the river. When prompted that the puzzle involves no cabbage, the next effort included taking a wolf across the river [1].

Why such preposterous responses? The reason lies in the functioning principle of the LLM. LLMs work by predicting the probability that a token (usually word) follows a string of tokens in a particular context [2]. The weird response is thus due to the close resemblance of the input puzzle to the old (and more difficult) puzzle of a farmer needing to take a wolf, a goat and a cabbage across the river. Since the training material is enormous – in the case of OpenAI's GPT-4, for example, it is said to be the entire Internet – the LLM will include the more difficult puzzle many times. Thus, detecting the tokens associated with man, goat, boat and river, the large language model predicts the next tokens to come from that puzzle. When the prompted puzzle is actually much simpler, the model is not able to detect that, given the scarcity of the simple puzzle in the training material. This is a good example of how the LLM functions: as a probabilistic model, it simply outputs the most likely (or one of the most likely) tokens associated with a string of tokens. It does not understand or reason in a human-like fashion, which is why it can be led astray so easily. [3,4]

When it comes to mathematical AI, the LLM-architectures thus seem inherently problematic. Mathematical deduction is not probabilistic. Instead of detecting patterns in data, a successful mathematical AI application has to follow rules – corresponding to the axioms and rules of proof in formal systems of mathematics. Against this background, it is hardly surprising that LLM-based AI systems have found relatively little use in mathematics. However, this may change when LLM-

architectures are combined with rule-based systems for new types of hybrid systems. In this kind of system, the LLM is used to generate potential solutions to mathematical problems, which are then tested on a rule-based system. Such a hybrid approach received a big boost in 2024, when Google's DeepMind reported significant success with its *AlphaGeometry* and *AlphaProof* applications in solving problems of the International Mathematical Olympiad [5,6]. Combining the *Gemini* large language model and the rule-based theorem prover *Lean*, the applications are exactly that kind of hybrid – sometimes called *neuro-symbolic* – AI.

In this paper, I investigate the potential of such AI tools in research mathematics, and their ethical consequences. While current theorem proving tools used by mathematicians are rule-based systems and as such have limited functionality, it is conceivable that a neuro-symbolic hybrid system could provide new proofs, including new mathematical theorems, autonomously. Such an *autonomous automated theorem prover* (AATP) could be transformative to mathematical practice. In the extreme case, new mathematical proofs could be generated simply by entering a system of axioms and rules of logic as input to an AATP. In such case, the human contribution would be minimal, yet it could lead to important achievements in the mathematical community. But even in less extreme cases, AI tools could be used to replace much of what is currently valued in the work of human mathematicians. This raises important questions about the future of mathematics. In this paper, I present epistemological considerations based on such a scenario. However, I will ultimately focus on the *ethics* of using such mathematical AI applications.

The paper is structured as follows. In Section 2, I present a short history of AI applications in theorem proving. In Section 3, I present scenarios of how this may change, based on AI architectures similar to AlphaProof and AlphaGeometry 2. From there, in Section 4, I move to the question of AI trust in the case of theorem proving applications. Finally, in Section 5, I examine what kind of ethical guidelines there should be for the use of AI tools in theorem proving. As the basis for my analysis, I use the guidelines presented for AI use in scientific research by Resnik and Hosseini [7]. However, I argue that the special characteristics of mathematical research requires the introduction of mathematics-specific instructions and regulations.

2. AI and theorem proving

Ironically, considering the weakness in mathematical tasks in celebrated modern-day AI systems like ChatGPT, in early days of AI research mathematics was considered to be one of the main success stories.[1] Indeed, one of the very first AI systems could prove mathematical theorems. The *Logic Theorist*

[1] Here I will use "artificial intelligence" to refer widely to any computer application designed to process tasks previously thought to require (human) intelligence. I acknowledge that the early developments no longer fit the definition of AI associated (either implicitly or explicitly) with much of the modern use of the term.

by Newell, Simon and Shaw [8] proved theorems of *Principia Mathematica* by Whitehead and Russell [9] with remarkable success: of the 52 theorems of the second chapter of the book, Logic Theorist managed to prove 38, in one case even providing a proof that was considered superior to that presented in the book [10].

Logic Theorist was a rule-based system, an early application of what is now called "good old-fashioned AI". Such systems have proven to be useful tools for mathematicians. They function based on following simple rules of logic, which make them reliable companions for humans solving mathematical tasks. The most common usage of such tools is for calculations, but they have also found success in theorem proving. Famously, computers have been used to prove conjectures like the four-color theorem [11] and Kepler's conjecture [12]. These were proofs by *exhaustion*, using the vast computational power of computers to one by one verify a finite number of cases. Shortly after Appel and Haken presented their proof, there were doubts about the reliability of this kind of method, given that we cannot check the calculations and the computer might make errors [13]. Nowadays, however, such worries are rarely heard. When we know which rules the computer is following, we trust it to do so reliably.

Although their reliability is no longer questioned (at least not in the same sense), proofs by exhaustion are somewhat frowned upon for their lack of mathematical elegance. In mathematics, in addition to proving a theorem, we would also like to understand *why* the theorem is true. Simply crunching the numbers case by case does not give us insight into that. In addition, proof by exhaustion works only in when the cases to be checked are finite when in mathematics theorems are typically proved for infinite domains.[2] Hence, while proofs by exhaustion provide the most famous cases of computer-assisted proofs in mathematics, they are not representative of the most common uses of AI tools in theorem proving. For that purpose, mathematicians use *automatic theorem provers* (ATP) or *interactive theorem provers* (ITP). Famous such software include *E, Vampire, Lean*, and *Mizar*.[3] What these software typically do is take a problem as the input, consisting of a set of first-order axioms and a conjecture (a first-order formula). Then, standardly using first-order logic with equality, the software checks whether the conjecture follows from the axioms [14]. If the conjecture holds, it is desirable that the ATP can provide a derivation of the theorem from the axioms, or in the case of ITP, parts of the derivation to help the human mathematician. From their functionality, it becomes clear why such software are often also called *proof assistants*: while they can be helpful tools for the human mathematicians, they do not provide proofs autonomously, making them more like pocket calculators than anything resembling *intelligent* applications [15].

Unsurprisingly, automated and interactive theorem provers are considered to be uncontroversial tools in mathematical practice, as the "real" work is done by the mathematician. There is nothing unethical in

[2] The great achievement in the cases of the four-color theorem and Kepler's conjecture was that the mathematicians were able to reduce the infinite domain to a finite number of cases to be checked.
[3] Often the same software has both functionalities.

using ATP or ITP software, and one is not expected to report their use in published papers, any more than one is expected to report, say, the usage of a spell-checker in correcting typos. They are simply modern tools at the disposal of the mathematicians: some find use for them while others do it never, but there is nothing controversial involved.[4]

However, this lack of controversy may well be due only to the limited functioning of the present-day software. Since even for the most software-savvy mathematician, the important mathematical content has to be generated by the human, the proof assistants are considered acceptable tools. But what if the functionality of the automatic theorem provers would be expanded? What if they could prove theorems essentially *autonomously*? In the case of such *autonomous automatic theorem provers* (AATP), serious questions concerning research ethics could arise. What if an AATP provided a proof of a *new* theorem autonomously?

For a long time, such AATP applications did not seem like a realistic prospect. Strictly speaking, of course, a classical ATP can be used to prove new theorems. By entering a system of axioms and rules of proof, we can simply have the ATP generate theorems of the system. The problem, however, is to limit this output to theorems and proofs that are somehow interesting to human mathematicians. While some progress has been made in terms of, for example, length of proofs [16–18], such measures do not help us distinguish between important and trivial *theorems*. Suggested criteria for interesting mathematics have included *insightfulness* [19,20] and *beauty* [21,22], but such notions remain vague and as such impossible to formalise for applications in automatic theorem provers (for a more detailed analysis, see [15]).

In this regard, however, the limitations only apply to rule-based systems. In [15], I have analysed how the matter could potentially change if we had machine-learning-based theorem provers applied for theorem proving. Importantly, such applications could detect patterns of what makes theorems humanly interesting without the patterns being fed explicitly as input. This kind of approach can give rise to a dual processing of theorems: a deep neural network being used to detect patterns and predict proof sequences, and a traditional rule-based theorem proving then used to test the proofs. When I submitted that paper, these types of *neural theorem provers* (NTP) were starting to be discussed among computer scientists [23–25] but they had not received much wider attention. This changed dramatically in July 2024 when Google's DeepMind published its results on their mathematical AI applications *AlphaProof* and *AlphaGeometry2*. These applications were reported to achieve new levels of artificial mathematical reasoning. Since they are likely to become standard applications to refer to in discussions concerning mathematical AI, I will focus on them in what follows.

[4] From personal communication with mathematicians, I have gotten the impression that the use of proof assistants is actually quite rare among research mathematicians. However, to the best of my knowledge, no reliable up-to-date data on this matter is available.

3. AlphaProof, AlphaGeometry 2, and the future of theorem proving AI

Google's DeepMind announced the AI model AlphaGeometry and its success in solving problems of the International Mathematical Olympiad (IMO) in January 2024 [5]. In July that year, it announced the model AlphaProof and a new version of their geometry model, AlphaGeometry 2. Also in that case, success in the IMO was reported [6]. While the AlphaGeometry models are restricted to geometrical problems, AlphaProof can solve problems also in other fields of mathematics. Since their training and functioning is very similar, I will focus here on DeepMind's report on AlphaProof.

AlphaProof is a machine learning system that is trained to prove mathematical statements in the programming language of the proof assistant Lean. Its functioning is based on a pre-trained large language model and the application of DeepMind's *AlphaZero* reinforcement learning algorithm (the same algorithm that famously reached very high levels in games like chess and go). A key feature of AlphaProof is that a *Gemini* large language model is fine-tuned to translate natural language statements in the training data into formal statements processable in Lean. In this way, the model was able to take roughly million informal mathematics problems and turn that into a database of 100 million formal problems. When presented with a problem, AlphaZero then uses this database to generate solution candidates, which are consequently processed in Lean as proof steps. Every successful proof step is then used to reinforce AlphaProof's large language model, improving its capacity to solve future problems [6].

It is important to note that AlphaProof is thus a hybrid of two approaches to AI. First, the Gemini large language model is trained with a deep neural network transformer architecture, which is the standard way modern machine learning systems work. Second, the solution candidates are processed in the rule-based, symbolic programming language of Lean. This kind of hybrid approach is sometimes called *neuro-symbolic AI* because it combines neural network architectures with symbolic, rule-based systems [26]. This idea is not new, it was supported prominently by Marcus in his book *The Algebraic Mind* [27]. It mirrors Kahneman's [28] dual "fast and slow" system theory of the mind, in which one system is responsible for "fast", unconscious and intuitive thinking while another system is responsible for "slow", conscious and deliberate thinking. This connection is explicitly mentioned by DeepMind in presenting AlphaGeometry:

> AlphaGeometry is a neuro-symbolic system made up of a neural language model and a symbolic deduction engine, which work together to find proofs for complex geometry theorems. Akin to the idea of "thinking, fast and slow", one system provides fast, "intuitive" ideas, and the other, more deliberate, rational decision-making. [5]

The working principle is thus that the large language model, being able to detect patterns in the training data, can predict successful constructs in a "fast" manner, which are then processed by the symbolic-

deductive engine in a "slow" manner. Essentially, the large language models associated with AlphaProof and AlphaGeometry are applied to find potential solutions to problems, which are then tested by the symbolic-deductive system.[5]

The performance of AlphaProof and AlphaGeometry has been impressive. In January 2024, it was reported that the original AlphaGeometry solved 25 out of the 30 geometry problems in the International Mathematical Olympiad (IMO), which is very close to the gold-medallist performance (25.9 on average) [5]. In July 2024, Alphaproof and AlphaGeometry 2 were reported to score 28 out of 42 points in the general IMO test. This is again very close to the gold-medallist performance level of 30 points [6].[6]

While at present these AI systems have been used to solve problems presented as parts of competitions, it is feasible that they can also be used to solve problems in actual research mathematics. This potential use can be divided into four categories:

1. They can be used to fill in "boring" parts of the proof, similarly to how ChatGPT is used by coders currently.

2. They can be used to find completely new proofs for known theorems. Some theorems, like the Four-color theorem, currently have proofs that many mathematicians consider inelegant or otherwise sub-optimal. The AI systems could find more satisfactory proofs.

3. They could be used to prove conjectures. Unproven conjectures, like Goldbach's conjecture, could be fed as the input to the system, with the task of finding a proof for it from a given system of axioms.

4. They could be used to prove completely new theorems, once that were not previously conjectures in the literature. In this scenario, the AI system generates a completely new theorem and a proof for it.

All four aspects would replace something that is currently the domain of human mathematicians. However, the first one differs in an important way from the other three. Namely, the first use is something that is already possible (to some degree) with the current generation of automated theorem provers. The

[5] In this, the functionality of AlphaProof and AlphaGeometry 2 differs from that of OpenAI o1, another AI application published in 2024 with reported success in solving mathematical tasks [29]. While o1, a development of the GPT-4o large language model, also uses reinforcement learning, it does not apply a rule-based system to check the results. OpenAI reports o1 as performing "complex reasoning" by using its "chains of thought – even going as far as declaring that it "thinks before it answers" [29] – but given that the answers are not tested on a rule-based system like in the case AlphaProof and AlphaGeometry 2, there is reason to doubt that the "reasoning" is *robust* in the sense that it can be relied upon generally. As of writing this, their success has not yet been properly tested, and they may well show to be bad at rare reasoning tasks – i.e., ones that are rare or absent in the training data – just like other LLMs have proven to be. See [4] for more.

[6] It should be noted that while AlphaGeometry solved the problems within the time limit, the AlphaProof and AlphaGeometry 2 combination sometimes took days to solve the problems, which is way above the 4.5 hours that the human participants have [6].

other three uses, as explained in the previous section, are not feasible with the current rule-based systems. For that reason, I will focus on them for the rest of this paper.

The first thing to note is that none of the last three features may be easy to achieve in practice. The success of AlphaProof, for example, is tightly connected to the developers being able to create the training dataset of 100 million formal problems. In the case of research mathematics, creating a comparably large training dataset provides a much more serious challenge. There do not exist millions of theorems and proofs, so the developers of a large language model for theorem proving would need to find a way to generate the dataset. This "scaling-up" is potentially problematic when we consider the requirement that the theorems should somehow be humanly *interesting*: after all, the generation process would essentially be about generating large sets of humanly interesting theorems that have not been considered humanly interesting [15].

Nevertheless, while the prospect of training a theorem proving AI with similar functioning to AlphaProof and AlphaGeometry 2 is undoubtedly a more difficult task, their success does give reason for optimism that it could be achieved. The training dataset generation could, for example, be based on applying a metric of "closeness" to proofs and theorems found in mathematical literature. That way, the dataset could remain sufficiently connected to actual human mathematics, thus potentially allowing the detection of patterns of humanly interesting mathematics. With such a development, a neuro-symbolic system could then become an autonomous automated theorem prover: we would simply need to present it with a system of axioms and rules of logic, and the system could generate humanly interesting mathematical theorems and their proofs. Instead of traditional proof assistants, such an application would be a proper AATP in the sense that human interaction with it would only concern the input and the output.

Training the large language model to identify patterns of humanly interesting mathematical content is a problem particularly in the fourth category, proving completely new theorems. An important part of mathematical practice is moving mathematics forward to interesting new directions. By proving new proofs, or parts of proofs, an AI system could certainly contribute in important ways to that. But to be comparable to human mathematicians more widely, an AATP would need to possess the capacity to identify entirely new problems in mathematics, ones that human mathematicians would ultimately agree on as being interesting. In terms of having enough training material to train the AI system, this aspect may be particularly difficult. Whereas patterns in the steps of humanly produced mathematical proofs may be easier to detect, it is not clear that classes of interesting mathematical problems – from a research mathematician's perspective – contain similar patterns. Given that the entire corpus of interesting problems in mathematical literature is nowhere near the kind of scale that is typically needed to train large language models – in the case of AlphaProof, for example, a million natural language problems were used to generate the formal dataset – it is far from trivial that a sufficiently large dataset for pre-training the model can be created. This is not to suggest, however, that progress could not be made. With

the help of criteria for similarity of problems, datasets could be developed that gradually become more useful for detecting patterns in what kind of mathematics is considered interesting.

To sum up, it is far from obvious that the AlphaProof/AlphaGeometry 2 approach can be scaled up to top-level research mathematics. However, neither are there any compelling reasons to think that no significant success could be achieved in that pursuit. A hybrid, neuro-symbolic, AI system could feasibly be developed based on the presently applied architectures that manages to prove new interesting theorems. Importantly, this could be done by an AATP that receives as the prompt only the task of finding a new theorem in a set of axioms (and rules of logic). In the final stage of the process, a generative AI tool could even write a mathematical paper based on the proof. For a long time, this step – producing a natural language article presenting the proof – would probably have been considered one of the most difficult, if not *the* most difficult obstacle. However, with generative AI applications based on large language models, this problem is already to large extent solved. ChatGPT, for example, has already been used to generate entire scientific papers [30]. While such papers are unlikely to have interesting scientific content, they are largely undistinguishable from humanly produced research articles when it comes to linguistic and structural aspects.

Therefore, as detailed above, the key question concerning the use of autonomous automated theorem provers is whether they can provide interesting *mathematical* content. AI tools can already be used in writing mathematical papers, as well as in checking the validity of mathematical proofs. In the scenario that I have presented in this paper, however, an ATTP would take the role of computers in mathematics to a new level. If this scenario becomes actualised, it will be transformative to mathematical practice. At that point, mathematicians will have access to computer tools that can potentially do much of what we have used to consider the exclusive domain of humans. In moderate applications, these tools can be used like ITP and ATP tools currently: they can be used to assist the human mathematician in producing and checking proofs. In the most extreme case, however, these tools can be used to essentially bypass the work previously demanding human mathematicians. An AATP could potentially generate a proof of an original theorem and present it as a scientific research article. In this latter scenario, we must ask (at least) two questions, often noted in the general literature of using AI tools in scientific research. First, can we *trust* the AI tools [31]. Second, what is *ethical use* of the AI tools [7]. In the next two sections, I will focus of these issues, respectively.

4. Trust in mathematical AI

In the philosophical discussions on artificial intelligence, trust in AI has been a key issue whether the topic is the reliability of AI technologies or their ethical use (see, e.g., [32,33]). When introducing new AI technologies, we are understandably concerned about their trustworthiness, along many dimensions. Can we trust the technology to function as described by its developers? Can we trust the users to apply

the technology in appropriate ways? What safety procedures are in place in case these expectations are violated? Where does the responsibility lie in such cases? Questions of this type are crucial in AI applications in high-responsibility areas like self-driving cars, medical AI technologies, or legal AI. But they are also relevant in the field of mathematical AI, including theorem-proving. Due to mathematical applications in technology, for example, the body of accepted mathematical knowledge has significance beyond the world of research mathematics. Hence, it is important that also mathematical AI is developed in a responsible way to ensure that we can trust in the applications.

Alvarado [34] has argued that AI tools are first and foremost *epistemic technologies*, i.e., they are designed and deployed for the particular purpose of expanding our capacities for knowing. Consequently, he has argued that *trust* in AI should be understood as *epistemic trust* and cannot be modelled after other forms of trust in technology, such as pharmaceuticals [31]. Here I apply Alvarado's approach for developing AI tools for mathematical research. If we trust a mathematical AI system, we trust it primarily as a tool for acquiring mathematical knowledge. Applications using this knowledge may be used for scientific purposes that necessitate wider ethical considerations concerning trust, but for present purposes I will limit my considerations to developing mathematical AI that we can trust as a reliable source of mathematical knowledge.

As mentioned in Section 1, trust in computer proofs was an important issue in the philosophy of mathematics when the first computer-assisted proofs emerged [11,13]. The concerns included, among other things, the possibility of malfunctioning computers and the inability to detect that. From a modern perspective, such concerns may seem somewhat odd. The possibility of bugs and hardware malfunctions has of course not disappeared, but what has changed is the understanding of *human* reliability in mathematics.[7] Nowadays, the idea that humans checking vast calculations could somehow be more reliable than a computer is likely to find little support.

That does not imply, however, that trust in mathematical AI is not an important topic to consider. Instead of questioning the reliability of rule-based systems, however, the main concern should be trust in machine learning systems. AlphaProof and AlphaGeometry 2 are both based on large language models and as such they potentially suffer from all the reliability problems associated with LLMs. This includes "hallucinations", i.e., the AI system generating incorrect outputs.[8] But perhaps the most commonly identified issue regarding trust in LLMs is the "black box" problem [37]. Even when deep neural network systems (like LLMs) are highly predictive, we rarely have a clear idea *why* they are so (see, e.g., Kay, 2018). Already due to the sheer complexity of the model, it is impossible to trace the algorithms that the system uses to come up with a particular output. In AI research, this problem is well-

[7] This is a topic in the philosophy of mathematical practice, for introduction see [35].
[8] "Hallucination," although the standard industry term, is a bad metaphor for this phenomenon: see [36].

known and it has given birth to the research field of *explainable AI* (XAI), which aims to find ways of making the processing of AI systems more transparent [39–41].

So far, XAI approaches have had limited success. The black box problem remains as important as ever, and any application using deep neural network architecture is vulnerable to it. This is also the case for AlphaProof and AlphaGeometry 2. It is not feasible to trace how the large language model is applied to get particular suggestions for proof steps. Neither is it feasible to trace how the reinforcement procedure changes the model. In this sense, both applications are as vulnerable to the black box problem as any deep-learning-based AI systems. However, in terms of the philosophical question of AI trust, there is an important difference to most other generative AI systems. A user of ChatGPT, for example, will receive a response to their prompt and needs to assess the reliability to the response. The response itself, however, may give relatively little information to help that assessment (in terms of sources, etc.). This is different with mathematical problem-solving applications like AlphaProof and AlphaGeometry 2. Instead of simply giving the correct solution as the output, they are trained to also provide the deductive steps that lead to the solution. Therein lies the strength of the neuro-symbolic systems: while the black box problem is no less serious in the large language model part of the system, the rule-based part of the hybrid system checks that the deductive steps are valid. Consequently, the hybrid system can give an explanation for the solution.

Granted, the explanation is not the kind of solution that XAI approaches are looking for in making AI explanations more transparent. The deductive steps presented by the neuro-symbolic AI system may not mirror in any way the processing of the LLM. However, in terms of AI trust in mathematics, this may not be a particularly serious issue. After all, in mathematical practice proofs are meant to be assessed based on their logical validity, not the thought process that led to them. Indeed, currently we know very little about the cognitive processes involved in research mathematics, yet few would consider this a problem. For this reason, I submit, the black box problem with regard to AI trust is not as damaging in the present scenario – i.e., neuro-symbolic theorem provers – as it is generally in generative AI. In the case of theorem-proving AI, we can – indeed, must – assess the reliability of the system based on the output, i.e., the deductive proof sequence. Assessing the reliability of a proof is of course not a trivial problem in mathematics. Automated proof checking is still limited, and proofs may be extremely long and resist successful checking by human mathematicians.[9] However, there is no reason to think that this problem is essentially different for human-generated and AI-generated proofs. Indeed, there is possibility that AI-generated proofs could become easier to check through automated procedures. A widespread use of some neuro-symbolic theorem prover application would also mean that the proofs

[9] A famous example of this is the proposed proof of the *abc* conjecture by Shinichi Mochizuki, which was so long and impenetrable that to this day it has remained in a kind of limbo of neither being confirmed or disproven (see, e.g., Ball, 2012).

would become more commensurable, applying the same programming language (like Lean in the case of AlphaProof). This can assist in developing tools for proof checking.

To sum up, in terms of AI trust, mathematics – or at least theorem proving – seems to be a special case. The practices of the mathematical community for assessing the reliability of mathematical content can generally be applied also for content generated by AI systems, including AATPs. This is not to say that adjustments would not need to be made. It remains to be seen what the AI-generated proofs would be like, but it is possible that they are generally longer than human-generated proofs. In this case, using human checking like in the present peer-review system may be more problematic in the case of AI-generated proofs. On the other hand, AI-generated proofs may be more suitable for automated proof checking, which can potentially make them more trustworthy than human-generated and human-checked mathematical proofs.

5. Ethics of using theorem proving AI

Issues of trust regards AI systems are closely connected to ethical issues like accountability and responsible use [43]. While mathematical AI has not received much attention in the philosophical literature, the use of AI tools in scientific research in general has been widely discussed in recent years [7]. The consensus among researchers seems to be that AI tools have come to scientific research to stay, whether as tools in education, assisting in research, or in scientific publishing. In all areas, there are multiple uses for AI tools. In scientific publishing, for example, AI tools can be used in various ways in the writing process, but also in the review and publication process. As the use of such tools becomes more prevalent, they are likely to change the entire publication process, perhaps to the level of disrupting it fundamentally [44].

The area of applying AI tools in scientific research that has been most comprehensively dealt with in present guidelines is using them in writing and editing articles. As reported in [7], documents like the European Code of Conduct for Research Integrity [45] and National Institutes of Health Guidelines [46] provide guidelines for using AI tools in writing and editing articles. As mentioned above, this is a relevant issue also for mathematical research. Hence guidelines specific to mathematics should be presented by publishers of journals and books. However, writing and editing articles is only a small part of mathematical and generally scientific research. Consequently, many have argued for introducing clear guidelines for the use of AI in scientific research and publishing [7,47].

I agree with this approach when it comes to introducing new theorem proving AI tools in mathematical research. As the starting point of my analysis of how mathematical research in particular should be understood in this respect, I will use the following list of six guidelines presented by Resnik and Hosseini [7]:

(1) Researchers are responsible for identifying, describing, reducing, and controlling AI-related biases and random errors

(2) Researchers should disclose, describe, and explain their use of AI in research, including its limitations, in language that can be understood by non-experts

(3) Researchers should engage with impacted communities, populations, and other stakeholders concerning the use of AI in research to obtain their advice and assistance and address their interests and concerns, such as issues related to bias

(4) Researchers who use synthetic data should (a) indicate which parts of the data are synthetic; (b) clearly label the synthetic data; (c) describe how the data were generated; and (d) explain how and why the data were used

(5) AI systems should not be named as authors, inventors, or copyright holders but their contributions to research should be disclosed and described

(6) Education and mentoring in responsible conduct of research should include discussion of ethical use of AI.

How should these guidelines be applied in the special case of mathematical research? Here I will evaluate each specifically for mathematics.

Guideline (1) is based on a well-known general problem of machine learning methodology. If the training data carries biases, these are likely to be present, or even amplified, in the outputs of the AI system. This problem has been discussed widely in the case of medical AI applications [48]. The training data may not represent the entire population, thus making subgroups – often disadvantaged populations – vulnerable for biased AI-based decisions. Similar concerns have been discussed in legal AI [49], among other fields. The other problem mentioned in the guideline, random errors, is also widely recognised. No method of data collecting is immune to errors, but in some medical applications, for example, AI machine learning methods have been associated with unusual amount of random errors [50].

How does Guideline (1) relate to theorem proving AI? In the kind of neuro-symbolic hybrid architecture that we focus on in this paper, the problem applies to the machine learning side, i.e., the pre-training of the large language model. Certainly, the datasets used for training the model can contain bias. This bias can be based on many visible factors: language, geography, publication status, etc. But inherent in the dataset of mathematical proofs can also be hidden biases based on gender, ethnicity, and other factors. In case of mathematical proofs, the bias may not be as prominent or damaging as in the medicinal or legal fields. However, also as a mathematical guideline, the choice of training data should be made transparent, and any biases should be minimised. The same goes for random errors, which are generally likely to be a lesser problem in the field of mathematics, due to the nature of the dataset. Mathematical

proofs may carry errors and gaps that cause errors in the output, but these are unlikely to be as common as in areas involving more error-prone procedures like image processing. More importantly, the hybrid nature of the AI applications discussed in this paper enable error detection and quality improvement in the reinforcement stage of the training procedure. Of course errors may still remain, so as a guideline it is important to minimise them also in the case of theorem proving AI.

Guideline (2) calls for explaining how AI was used in the research in an open fashion, in language understandable to non-specialists. This is clearly something that should be applied in the fields of mathematics, as well. Indeed, I submit that this guideline should have been introduced to mathematical research already. At present, it is not required to disclose the use of AI tools for checking proofs, or even generating parts of proofs. This is the case, for example, with the guidelines published by the American Institute of Mathematical Sciences. [51].[10] While the use of generative AI, such as ChatGPT, is required to be mentioned under those guidelines, they do not mention anything about using AI tools for checking mathematical content. This may become problematic as the line between using AI tools for checking proofs and generating new content can become blurred. It would be better for transparency to require disclosing the use of AI tools in any stage of processing the mathematical content. In addition, using AI tools for generating text should always be disclosed, and any AI-generated content should be carefully checked.

The main problem with Guideline (3) is the difficulty of determining what communities, populations, or stakeholders are impacted by mathematical AI research. Given the wide range of applications of mathematics in science, but also wider in society, an argument could be made that the impact of mathematics reaches almost everywhere. Even if we limit ourselves to, say, educational contexts, the scope remains unrealistically wide for proper engaging. Thus, I suggest that the suitable scope for engagement is that of research mathematicians. With this limitation, this guideline can be directly applied to theorem proving AI. Research using such tools should engage with the research mathematician community to disclose and discuss the impact and potential problems.

With regard to Guideline (4), we can replace "synthetic data" with "AI-generated proofs". These may involve AI use of any of the four types disclosed in Section 3, but in each case, it should be disclosed what type of AI-generated content was produced. In the first case, for example, the parts proven by the AI should be clearly marked in the proof. In the second and third cases, it should be openly disclosed that the proof was generated by an AI system. Finally, in the fourth case, it should be transparent that also the proven conjecture was generated by an AI system. The AI application(s) in question should be identified and their most important characteristics explained. In addition, the reason for using them should be made clear, as well as the researchers' role in the process.

[10] There generative AI is distinguished from "assistive AI" (such as spelling checker functionality in Microsoft Office), whose use is not required to be reported.

The Guideline (5) is often included in guidelines for AI use (see, e.g., [51]). The motivation for it is that the human authors should retain the entire responsibility for the published research. In the case of AATPs, this may ultimately become the main role of the human author. If an ATTP can autonomously generate a proof of a new theorem and produce a scientific paper presenting it, what role remains for humans in the process? In order to retain human accountability for published research findings, even in cases of minimal human contribution, human authorship is required. Aside from being accountable for the publication, the role of the author would also be to explain aspects of the publication when needed.

The Guideline (5) calls for the AI contributions to be "disclosed and described". I agree with this, but I believe that we should set clear guidelines for maximally visible acknowledgments of AI-generated proofs. For this, in the second and third types of AI use presented in Section 3, my suggestion is that the title of the paper should include the words "an AI-generated proof". For the fourth type, the words should be "an AI-generated theorem and proof". For the first type, the matter depends on the extent of AI use. If the use is not central, it would be enough to acknowledge this in the article. But for more extensive use of AI tools, the title should contain the words "an AI-assisted proof". The specifics of what counts as extensive use need to be determined by mathematicians.

The Guideline (6) is directly applicable to research mathematics. The ethical use of AI, including guidelines like the ones presented here, need to be part of mathematical education and mentoring in universities.

Equally important as setting guidelines for the proper use of generative AI tools in mathematical research is to agree on protocols for cases in which those guidelines are broken. This topic is the divided into two. First, there is the question of *detecting* misuse. Second, there is the question of determining the *consequences* for misuse. Both issues are very difficult. The need for reliable tools for detecting AI-generated text was recognised quickly after chatbots like ChatGPT were launched (see, e.g, [52]), yet relatively little has been achieved in the field. Tools have been developed and AI-detection is likely to become increasingly big business, but in terms of detecting AI-generated content with sufficient reliability to impose consequences on authors, important problems remain.[11] This also makes determining the consequences difficult, as potentially career- and life-changing sanctions would need to be imposed based on insufficient evidence. Due to such fundamental problems, policies for detection and disciplinary actions may be very hard to introduce in practice. This further underscores the importance of agreeing on ethical guidelines in a timely fashion: establishing fair and ethical research practices for AI tools can help deal with potential misuse when their application becomes more widespread.

[11] To give one example, OpenAI's tool for detecting AI-generated text was reported in January 2023 to identify correctly only 26% of AI-generated text (true positives) as being "likely AI-written" while it incorrectly labels human-written text as AI-generated 9% of the time (false positives) [53]. It should clear that such a low rate of true positives and high rate of false positives prevents any reliable use of the tool.

Conclusion

In this paper, I have argued that, just like the use of AI tools in scientific research in general, the use of AI tools in mathematical research requires ethical guidelines. Due to the special characteristics of mathematical practice, we cannot directly apply general guidelines, like those presented by Resnik and Hosseini [7]. However, those guidelines can be used as the basis for developing mathematics-specific instructions.

At present, AI tools based on machine learning systems do not play a major role in mathematics compared to many other fields of scientific research, like biology and chemistry. However, the early success of AlphaProof and AlphaGeometry 2 suggests that mathematics may be impacted equally or perhaps even more profoundly by the development of AI. In this paper, I have presented the scenario of an autonomous automated theorem prover (AATP), which can generate proofs for new mathematical theorems with minimal human contribution. When it comes to mathematical practice, such tools are potentially transformative. In that scenario, mathematical theorem proving may become more about the skilful use of AI tools than about the traditional skills and knowledge connected with mathematics.

Hence, a lot is at stake for mathematical communities. In the worst case, publication records and consequently even careers in mathematics could be created by being a skilful user of AI. If AI use is neither declared nor detected, dishonest people can use AI tools to create a misleading image of expertise. Indeed, this might be happening already, even though AATPs have not yet been introduced. In educational settings, using AI tools for problem solving is a very problematic prospect, and tools like AlphaProof could easily be used for that purpose. But as I have argued, the problem extends also to research mathematics, including proving new theorems. General guidelines and regulations for the ethical use of such applications – and hopefully also detection tools – should be established quickly.

Declaration of interest

The author declares no competing interests.

Declaration of funding

The author declares no funding.

References

[1] G. Marcus, This one important fact about current AI explains almost everything, Marcus on AI (2024). https://garymarcus.substack.com/p/this-one-important-fact-about-current (accessed September 25, 2024).

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020: pp. 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (accessed June 11, 2024).
[3] N. Yax, H. Anlló, S. Palminteri, Studying and improving reasoning in humans and machines, Commun Psychol 2 (2024) 1–16. https://doi.org/10.1038/s44271-024-00091-8.
[4] R.T. McCoy, S. Yao, D. Friedman, M.D. Hardy, T.L. Griffiths, Embers of autoregression show how large language models are shaped by the problem they are trained to solve, Proceedings of the National Academy of Sciences 121 (2024) e2322420121. https://doi.org/10.1073/pnas.2322420121.
[5] DeepMind, AlphaGeometry: An Olympiad-level AI system for geometry, Google DeepMind (2024). https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/ (accessed October 10, 2024).
[6] DeepMind, AI achieves silver-medal standard solving International Mathematical Olympiad problems, Google DeepMind (2024). https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/ (accessed October 10, 2024).
[7] D.B. Resnik, M. Hosseini, The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool, AI Ethics (2024). https://doi.org/10.1007/s43681-024-00493-8.
[8] A. Newell, J.C. Shaw, H.A. Simon, Empirical explorations of the logic theory machine: a case study in heuristic, in: Papers Presented at the February 26-28, 1957, Western Joint Computer Conference: Techniques for Reliability, 1957: pp. 218–230.
[9] A.N. Whitehead, B. Russell, Principia Mathematica - Volumes 1-3, Cambridge University Press, Cambridge, 1910.
[10] P. McCorduck, C. Cfe, Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence, 2nd edition, A K Peters/CRC Press, 2004.
[11] K. Appel, W. Haken, Every planar map is four colorable, Bulletin of the American Mathematical Society 82 (1976) 711–712.
[12] T. Hales, M. Adams, G. Bauer, T.D. Dang, J. Harrison, L.T. Hoang, C. Kaliszyk, V. Magron, S. Mclaughlin, T.T. Nguyen, Q.T. Nguyen, T. Nipkow, S. Obua, J. Pleso, J. Rute, A. Solovyev, T.H.A. Ta, N.T. Tran, T.D. Trieu, J. Urban, K. Vu, R. Zumkeller, A FORMAL PROOF OF THE KEPLER CONJECTURE, Forum of Mathematics, Pi 5 (2017). https://doi.org/10.1017/fmp.2017.1.
[13] T. Tymoczko, The Four-Color Problem and Its Philosophical Significance, The Journal of Philosophy 76 (1979) 57–83. https://doi.org/10.2307/2025976.
[14] A. Voronkov, Automated Reasoning: Past Story and New Trends, in: IJCAI, 2003.
[15] M. Pantsar, Theorem proving in artificial neural networks: new frontiers in mathematical AI, Euro Jnl Phil Sci 14 (2024) 4. https://doi.org/10.1007/s13194-024-00569-6.
[16] B. Fitelson, L. Wos, Finding Missing Proofs with Automated Reasoning, Studia Logica: An International Journal for Symbolic Logic 68 (2001) 329–356.
[17] M. Kinyon, Proof simplification and automated theorem proving, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 377 (2019) 20180034. https://doi.org/10.1098/rsta.2018.0034.
[18] R. Veroff, Finding Shortest Proofs: An Application of Linked Inference Rules, Journal of Automated Reasoning 27 (2001) 123–139.
[19] D. Macbeth, Proof and Understanding in Mathematical Practice, Philosophia Scientiæ. Travaux d'histoire et de Philosophie Des Sciences (2012) 29–54. https://doi.org/10.4000/philosophiascientiae.712.
[20] K. Weber, Proofs that develop insight, For the Learning of Mathematics 30 (2010) 32–36.

[21] S.G.B. Johnson, S. Steinerberger, Intuitions about mathematical beauty: A case study in the aesthetic experience of ideas, Cognition 189 (2019) 242–259. https://doi.org/10.1016/j.cognition.2019.04.008.
[22] G.-C. Rota, The Phenomenology of Mathematical Beauty, Synthese 111 (1997) 171–182.
[23] D. Jenson, Automated Theorem Proving with Graph Neural Networks, Stanford CS224W GraphML Tutorials (2023). https://medium.com/stanford-cs224w/automated-theorem-proving-with-graph-neural-networks-49c091024f81 (accessed June 5, 2024).
[24] G. Lample, T. Lacroix, M.-A. Lachaux, A. Rodriguez, A. Hayat, T. Lavril, G. Ebner, X. Martinet, HyperTree Proof Search for Neural Theorem Proving, Advances in Neural Information Processing Systems 35 (2022) 26337–26349.
[25] H. Wang, H. Xin, C. Zheng, L. Li, Z. Liu, Q. Cao, Y. Huang, J. Xiong, H. Shi, E. Xie, J. Yin, Z. Li, H. Liao, X. Liang, LEGO-Prover: Neural Theorem Proving with Growing Libraries, (2023). https://doi.org/10.48550/arXiv.2310.00656.
[26] A. Sheth, K. Roy, M. Gaur, Neurosymbolic AI -- Why, What, and How, (2023). https://doi.org/10.48550/arXiv.2305.00813.
[27] G.F. Marcus, The Algebraic Mind: Integrating Connectionism and Cognitive Science, American First edition, Bradford Books, Cambridge, Mass., 2001.
[28] D. Kahneman, Thinking, Fast and Slow, 1st edition, Farrar, Straus and Giroux, New York, 2011.
[29] OpenAI, Learning to Reason with LLMs, (2024). https://openai.com/index/learning-to-reason-with-llms/ (accessed October 21, 2024).
[30] G. Conroy, Scientists used ChatGPT to generate an entire paper from scratch — but is it any good?, Nature 619 (2023) 443–444. https://doi.org/10.1038/d41586-023-02218-z.
[31] R. Alvarado, What kind of trust does AI deserve, if any?, AI Ethics 3 (2023) 1169–1183. https://doi.org/10.1007/s43681-022-00224-x.
[32] H. Choung, P. David, A. Ross, Trust in AI and Its Role in the Acceptance of AI Technologies, International Journal of Human–Computer Interaction 39 (2023) 1727–1739. https://doi.org/10.1080/10447318.2022.2050543.
[33] M. Ryan, In AI We Trust: Ethics, Artificial Intelligence, and Reliability, Sci Eng Ethics 26 (2020) 2749–2767. https://doi.org/10.1007/s11948-020-00228-y.
[34] R. Alvarado, AI as an Epistemic Technology, Sci Eng Ethics 29 (2023) 32. https://doi.org/10.1007/s11948-023-00451-3.
[35] P. Mancosu, ed., The Philosophy of Mathematical Practice, 1st edition, Oxford University Press, Oxford ; New York, 2008.
[36] M. Pantsar, R.E. Fabry, How Not to Talk about Chatbot Mistakes, (2024). https://philsci-archive.pitt.edu/23878/ (accessed October 16, 2024).
[37] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 4th edition, Pearson, Hoboken, 2020.
[38] K.N. Kay, Principles for models of neural information processing, NeuroImage 180 (2018) 101–109. https://doi.org/10.1016/j.neuroimage.2017.08.016.
[39] D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, arXiv Preprint arXiv:1710.00794 (2017).
[40] A. Holzinger, From machine learning to explainable AI, in: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), IEEE, 2018: pp. 55–66.
[41] J.A.F. Thompson, Forms of explanation and understanding for neuroscience and artificial intelligence, (2021). https://doi.org/10.31234/osf.io/5g3pn.
[42] P. Ball, Proof claimed for deep connection between primes, Nature (2012). https://doi.org/10.1038/nature.2012.11378.
[43] C. Novelli, M. Taddeo, L. Floridi, Accountability in artificial intelligence: what it is and how it works, AI & Soc 39 (2024) 1871–1882. https://doi.org/10.1007/s00146-023-01635-y.
[44] G. Conroy, How ChatGPT and other AI tools could disrupt scientific publishing, Nature 622 (2023) 234–236. https://doi.org/10.1038/d41586-023-03144-w.

[45] All European Academies, The European Code of Conduct for Research Integrity, Revised Edition 2023, (2023). https:// allea. org/ code- of- condu ct/.
[46] National Institutes of Health, Guidelines for the Conduct of Research in the Intramural Program of the NIH, (2023). https:// oir. nih. gov/ system/ files/ media/ file/ 2023- 11/ guide lines- condu ct_ resea rch. pdf.
[47] M. Hosseini, S.P.J.M. Horbach, Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review, Res Integr Peer Rev 8 (2023) 4. https://doi.org/10.1186/s41073-023-00133-5.
[48] M. Mittermaier, M.M. Raza, J.C. Kvedar, Bias in AI-based models for medical applications: challenges and mitigation strategies, Npj Digit. Med. 6 (2023) 1–3. https://doi.org/10.1038/s41746-023-00858-z.
[49] R. Rodrigues, Legal and human rights issues of AI: Gaps, challenges and vulnerabilities, Journal of Responsible Technology 4 (2020) 100005. https://doi.org/10.1016/j.jrt.2020.100005.
[50] P. Ball, Is AI leading to a reproducibility crisis in science?, Nature 624 (2023) 22–25. https://doi.org/10.1038/d41586-023-03817-6.
[51] American Institute of Mathematical Sciences, Guidelines for the Use of AI Tools in Writing and Research, (n.d.). https://www.aimsciences.org/index/GuidelinesforAI (accessed October 16, 2024).
[52] M. Heikkilä, How to spot AI-generated text, MIT Technology Review (2022). https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/ (accessed October 15, 2024).
[53] OpenAI, New AI classifier for indicating AI-written text, (2023). https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/ (accessed October 15, 2024).