HUMAN-MACHINE COLLABORATION IN MATHEMATICAL RESEARCH AFTER AGI

Anonymous authors
Paper under double-blind review

ABSTRACT

The prospect of artificial general intelligence raises concerns about the future role of human creativity in scientific research. In this paper, we argue that increasingly autonomous learning and reasoning systems are more likely to amplify human mathematical creativity than to replace it. We examine how such systems may differentially influence mathematical subfields, identifying areas where automation is likely to be highly effective, areas where human–machine collaboration will dominate, and areas that are expected to remain largely human-driven. Our analysis focuses on structural features of mathematical practice, including formalization, abstraction, and dependence on empirical experimentation. By mapping these features across major domains of mathematics, we provide a framework for understanding how advanced learning systems may reshape mathematical research in a post-AGI scientific landscape.

1 INTRODUCTION

Artificial Intelligence (AI) has revolutionized many aspects of the modern world, ranging from government (Margetts & Dorobantu, 2019), healthcare (Topol, 2019), and education (Holmes et al., 2019) to weather prediction (Dueben & Bauer, 2018) and space exploration (Erwei et al., 2025). Artificial General Intelligence (AGI) remains a topic of active philosophical and scientific debate, referring to the hypothetical ability of artificial systems to perform a broad range of intellectual tasks at a level comparable to or exceeding that of humans (Goertzel, 2014). The timeline for the emergence of AGI remains highly uncertain, even among leading AI researchers (Grace et al., 2018). While some estimates suggest it may emerge within decades, others argue that it may require a century or may never be realized at all.

Large Language Models (LLMs) constitute a foundational component of many modern AI systems. They are neural networks trained on vast corpora of text to predict the next token in a sequence, giving rise to emergent language understanding and reasoning abilities (Brown et al., 2020). Through layered attention mechanisms, LLMs learn internal representations that support coherent generation and structured reasoning across tasks (Vaswani et al., 2017). Recent advances have extended LLMs to multimodal reasoning, enabling them to process text, images, audio, and video for richer, context-sensitive interactions (Alayrac et al., 2022). Notable multimodal models include Gemini 2.0 (Google DeepMind), Molmo (Allen Institute for AI), and GPT-4o (OpenAI). Research on neural scaling laws indicates that increasing model size and data generally improves performance up to regimes of diminishing returns (Kaplan et al., 2020; Hoffmann et al., 2022). As a result, state-of-the-art LLMs now typically consist of billions of trainable parameters.

More recently, LLM-based systems augmented with reinforcement learning and formal reasoning tools have demonstrated promising results in mathematical discovery and automated proof generation (Polu & Sutskever, 2020). Notable examples include AlphaProof (DeepMind, 2024), Axiom-Math, and Gauss, which leverage formal proof assistants such as Lean. Despite these advances, current AI systems still lack the general intelligence exhibited by even relatively simple biological organisms in many real-world settings (Lake et al., 2017). Human intelligence, however, surpasses that of such organisms by large margins in abstraction and symbolic reasoning. This observation suggests that achieving AGI may require orders-of-magnitude scaling, novel architectural designs, or fundamentally different computational paradigms beyond current LLM-based approaches (Ben-

gio et al., 2021). In this paper, we examine structural criteria for predicting how the emergence of AGI may differentially impact various areas of mathematical research.

2 Impact of AGI on Mathematical Subfields

Mathematics, as a scientific discipline, requires immense ingenuity and profound human creativity, expressed through symbolic languages and formal reasoning. Since its early development, mathematical progress has been driven primarily by human intellectual contributions, often demanding sustained effort over long periods of time. Mathematicians may spend hours or even decades attempting to prove theorems or resolve fundamental problems. Some advances build incrementally on existing work, while others arise from novel insights inspired by interactions with nature or other scientific domains (Hadamard, 1945; Poincaré, 1908).

As artificial systems advance toward increasingly general forms of intelligence, it is reasonable to expect that their influence on mathematical research will be substantial. The goal of this work is not to speculate on the precise existence or timeline of Artificial General Intelligence (AGI), but rather to provide a consistent framework for understanding how such systems may impact different areas of mathematics, including category theory, combinatorics, logic, and partial differential equations (PDEs). Our framework is based on three key structural factors, which we describe below.

Abstraction. Abstraction refers to the extent to which a mathematical field operates on highly general concepts that are far removed from concrete objects, explicit computations, or physical interpretations. Highly abstract areas emphasize conceptual unification, the invention of new definitions, and reasoning about structures that may lack intuitive or empirical grounding (Mac Lane, 1998). Domains such as category theory, homotopy theory, and parts of the foundations of mathematics often progress through the development of new languages and perspectives rather than through direct manipulation of existing formal systems. In these settings, mathematical creativity manifests as the synthesis of disparate ideas and the construction of novel conceptual frameworks, relying heavily on human judgment, aesthetic intuition, and long-term vision which are qualities that are difficult to reduce to optimization objectives or pattern extraction.

In contrast, areas with lower levels of abstraction, such as finite combinatorics or graph theory, tend to operate on concrete objects with explicit representations (Diestel, 2017). These domains often admit algorithmic exploration, exhaustive search, and pattern discovery, making them more amenable to automation by advanced learning systems.

From an AGI perspective, abstraction acts as a natural barrier to domination. Though AGI systems may assist in navigating complex abstractions by suggesting analogies, checking consistency, or recalling related constructions, the creation of genuinely new abstract frameworks is likely to remain a collaborative human–machine endeavor. Thus, the higher the level of abstraction in a mathematical subfield, the more AGI is expected to amplify rather than replace human creativity.

Formalization. Formalization measures the extent to which mathematical reasoning in a given field can be expressed precisely within well-defined logical systems, symbolic languages, or proof assistants. Highly formalized fields allow statements, proofs, and transformations to be encoded explicitly as sequences of syntactic rules. Areas such as mathematical logic, proof theory, and parts of algebra and combinatorics exhibit a high degree of formalization (Avigad et al., 2014). In these domains, correctness is determined by strict adherence to axioms and inference rules, making them particularly suitable for automated theorem proving and reinforcement-learning-based reasoning systems. Indeed, recent advances in formal proof assistants demonstrate that machines can already match or exceed human performance in certain narrowly defined formal tasks.

However, formalization is often unevenly distributed even within a single field. For example, while parts of number theory can be highly formalized, the early stages of conjecture formation often rely on informal reasoning, heuristic insight, and analogical thinking. These informal stages resist full mechanization despite eventual formal encoding.

AGI is expected to have its strongest impact in regions of mathematics where formalization is both deep and mature. In such areas, AGI systems may take over routine proof construction, verification,

and exploration of large formal spaces. Importantly, this does not diminish human creativity; instead, it shifts human effort toward higher-level tasks such as selecting meaningful problems, designing new axioms, and interpreting the significance of results.

Experimental Dependence. Experimental dependence describes the extent to which a mathematical field relies on computational experiments, simulations, or large-scale data exploration to guide intuition, generate conjectures, or validate hypotheses prior to formal proof. Fields such as experimental number theory, random matrix theory, dynamical systems, and parts of PDEs often make extensive use of numerical simulations and empirical observation (Borwein & Bailey, 2008). In these areas, mathematicians routinely explore vast parameter spaces, observe recurring patterns, and use computational evidence to formulate conjectures.

AGI systems are particularly well suited to this mode of mathematical practice (Davies et al., 2021). Their ability to process large volumes of data, detect subtle regularities, and explore high-dimensional spaces enables them to dramatically accelerate the experimental phase of research. Rather than replacing mathematicians, AGI acts as a powerful exploratory partner by generating hypotheses at a scale and speed unattainable by humans alone.

Conversely, fields with minimal experimental dependence, such as axiomatic set theory or foundational logic, offer fewer opportunities for data-driven discovery. Progress in these domains depends primarily on philosophical considerations, internal consistency, and conceptual clarity, limiting the direct influence of experimental AGI systems.

Appendix B presents the mapping of these features across multiple mathematical domains, along with their corresponding levels of AGI impact. The mapping reflects how the fields are practiced today, not value judgments. For instance, statistics (MSC 62) is mapped to the structural features based on how the field is commonly practiced in modern research. Its level of abstraction is relatively low, as statistical methods are primarily grounded in concrete data, observable quantities, and empirically motivated models rather than highly abstract conceptual frameworks. Formalization is moderate: while statistics is built on rigorous mathematical foundations in probability theory and optimization, most research and applied work relies on informal proofs, asymptotic arguments, and empirical validation rather than fully formalized reasoning in proof assistants. In contrast, statistics exhibits very high experimental dependence, with progress driven by large-scale computation, simulation, and data analysis, including Monte Carlo methods, resampling techniques, and empirical benchmarking. Because AGI systems are particularly well suited to large-scale experimentation, pattern discovery, and automated model optimization, these structural characteristics imply a very high level of AGI impact.

3 Conclusion

Human intelligence is one of the most complex phenomena in the universe ((Dehaene, 2020). We view it as self-sustaining rather than self-annihilating, suggesting that advanced artificial systems are unlikely to subdue or eliminate it. Instead, greater emphasis should be placed on understanding how AGI operates internally and on developing methods to maximize its potential for the benefit of humanity.

References

Helen Margetts and Cosmina Dorobantu. Rethink government with AI. *Nature*, 568:163–165, 2019. https://www.nature.com/articles/d41586-019-01099-5

Eric J. Topol. *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books; ISBN: 9781541644632, 2019.

Wayne Holmes, Maya Bialik, and Charles Fadel. *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign, 2019. https://curriculumredesign.org/our-work/artificial-intelligence-in-education/

Peter D. Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11:3999–4009, 2018. https://gmd.copernicus.org/articles/11/3999/2018/

Xin-Yu Xiao, Yalei Liu, Xiangyu Liu, Zengrui Li, Erwei Yin, and Qianchen Xia. Lunar twins: We choose to go to the moon with large language models. *Findings of the Association for Computational Linguistics: ACL 2025*, 1325–1339, 2025.

Ben Goertzel. Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, 2014.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018. https://www.jair.org/index.php/jair/article/view/11222

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 2022.

Jared D. Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. https://arxiv.org/abs/2001.08361

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. https://arxiv.org/abs/2203.15556

Stanislaw Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020. https://arxiv.org/abs/2009.03393

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. https://doi.org/10.1017/S0140525X16001837

Yoshua Bengio, Yann LeCun, and Geoffrey Hinton. Deep learning for AI. *Communications of the ACM*, 64(7):58–65, 2021. https://dl.acm.org/doi/10.1145/3448250

Jacques Hadamard. *The mathematician's mind: The psychology of invention in the mathematical field*. Princeton University Press, ISBN: 9780691029313, 1996. https://press.princeton.edu/books/paperback/9780691029313/the-mathematicians-mind

Henri Poincaré. *Science and method*. Thomas Nelson and Sons, 1908. https://www.amazon.com/Science-Method-Henri-Poincare/dp/1602064482

Saunders Mac Lane. *Categories for the working mathematician*. Springer, 1998. https://link.springer.com/book/10.1007/978-1-4757-4721-8

Reinhard Diestel. *Graph theory*. Springer, 2017. https://link.springer.com/book/10.1007/978-3-662-53622-3

Jeremy Avigad, Leonardo de Moura, and Soonho Kong. Theorem proving in Lean. *Carnegie Mellon University.*, 2021.

Jonathan M. Borwein and David H. Bailey. *Mathematics by experiment*. A K Peters, 2008. https://www.amazon.com/Mathematics-Experiment-Plausible-Reasoning-Century/dp/1568814429

Alex Davies, Petar Veličković, Laurent Buesing, et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600:70–74, 2021. https://www.nature.com/articles/s41586-021-04086-x

Stanislas Dehaene. *How we learn*. Viking, 2020. https://www.amazon.com/How-We-Learn/dp/0241366461

A LLM USAGE STATEMENT

In the course of developing this manuscript, large language models were utilized for grammatical refinement and selective text revision.

B MAPPING THE FEATURES

This table shows the relationship between the three structural factors and different mathematical fields based on MSC2020-Mathematics Subject Classification Systems.

Table 1. Mapping of MSC2020 domains to structural factors and expected AGI impact

MSC2020 domain	**Abs.**	**Form.**	**Exp.**	**AGI**
01 History and biography	L	L	L	VL
03 Mathematical logic and foundations	M	H	L	H
05 Combinatorics	LM	H	M	H
08 General algebraic systems	M	H	L	MH
11 Number theory	M	MH	M	MH
13 Commutative algebra	MH	H	L	M
14 Algebraic geometry	H	M	L	M
15 Linear and multilinear algebra; matrix theory	L	H	M	H
17 Nonassociative rings and algebras	M	M	L	M
18 Category theory; homological algebra	VH	LM	L	L
33 Special functions	L	H	M	H
35 Partial differential equations	M	M	H	H
37 Dynamical systems and ergodic theory	M	M	H	H
43 Abstract harmonic analysis	H	M	L	LM
44 Integral transforms, operational calculus	LM	H	M	H
46 Functional analysis	H	M	L	LM
47 Operator theory	H	M	L	LM
49 Calculus of variations and optimal control; optimization	M	M	H	H
51 Geometry	M	M	L	M
52 Convex and discrete geometry	M	H	M	H
53 Differential geometry	H	M	L	LM
55 Algebraic topology	VH	M	L	L
60 Probability theory and stochastic processes	M	M	H	H
62 Statistics	L	M	VH	VH
65 Numerical analysis	L	H	VH	VH
76 Fluid mechanics	M	M	VH	H
90 Operations research, mathematical programming	L	MH	VH	VH

Legend: Abs.=abstraction, Form.=formalization, Exp.=experimental dependence, AGI=expected AGI impact. Levels: VL=very low, L=low, LM=low–medium, M=medium, MH=medium–high, H=high, VH=very high.