ATPs as Universal AIs: What Do AGI Architectures Suggest for ATP Research?

Zarathustra Amadeus Goertzel

Czech Technical University in Prague, Czech Republic

I propose to discuss the realization that automated theorem provers (ATPs) are universal artificial intelligence (AI) systems when using complete search strategies [38], and how considering ATPs as a form of artificial general intelligence (AGI) can suggest fruitful, necessary research directions. Furthermore, the AGI field may learn from the work that has gone into making complete proof searches efficient and integrating AI and ML techniques into the symbolic theorem provers.

Universal search [6] procedures work by cleverly enumerating all possible solutions (programs) increasing in size until a solution is found. Levin search solves *inversion problems* where, given a function f and a value y, the goal is to find a program that outputs an x such that $f(x) = y$. Hutter search solves for well-defined functions from a domain X to a domain Y, only working with functions that are provably equivalent to the target function with provable time-bounds, and works by enumerating proofs to find the programs to feed to the Levin search. With the work of Bentkamp et al. [1], we have an efficient refutation-complete superposition calculus for full higher-order logic (HOL): from any appropriately quantifier-normalized initial clause set, a refutation can be derived.

The scope of problems that can be formulated as HO refutation problems is very large: standard reinforcement learning problems used for training neural networks should be included. In a *constructive* proof-system, existential witnesses can be extracted from the proof. Performance constraints on the solutions should also be encodable. Universal search procedures all run into efficiency challenges: the Solomonoff induction-based reinforcement learning agent AIXI [14] is infamously incomputable. While a Monte-Carlo Tree Search AIXI approximation learned to play Pac-Man [34], one can argue that the ATP field constitutes the most developed approach to improving the performance of universal, general AI systems.

"No Free Lunch"-style theorems [23, 24, 27, 39] suggest that all effective, real-world AIs will need to engage in a dialectic dance between specialization and generalization. To this end, many AGI architectures [8, 20–22, 29, 35] have been developed that aim to specify functionalities needed for general functioning in a range of environments that humans care about and how these components should be connected. Premise selection is a motivating example in the ATP domain: theoretically, if it's mutually consistent, the whole Mizar Mathematical Library (MML) could be loaded into each proof search: given clause selection strategies and AI will determine whether theory clauses are relevant or not on the fly. Yet practically, isolating the *premise selection* module *between proof search runs* is essential to obtaining good performance.

Let's review the cognitive architecture of E with ENIGMA [11, 15]. Externally, E receives some command-line arguments, strategies, and then theory and problem descriptions, usually in a TPTP format [30]; E returns output on the results of the proof search. Internally, E explores a mathematical space defined by the initial clause set. E perceives this space via clauses (as term trees) and featurized vectors (for the AI). The goal is the empty clause, implicitly aiming for proofs with smaller terms. The cognitive cycle is the given clause loop: evaluate, select, generate, and simplify until saturated. Planning is done via term ordering, literal selection, symbolic time and weight-based strategies, and then AI for clause selection and filtering via gradient-boosted decision trees and/or graph neural networks (GNN). Thought-action is taken via the superposition calculus, which primarily amounts to resolution and term-rewriting. For short-term memory, E has the (un)processed clause sets, some of which can be fed to the GNN. For long-term memory, E has proof search data (episodic memory), proof vectors in ProofWatch [10], and AI models in ENIGMA (semantic & procedural memory). These are the core components of a "traditional" cognitive architecture!

What features are missing that many AGI architectures contain? One debatable feature is *autonomy* [32], which may only be needed for AGI systems in certain domains (such as controlling a Mars rover or game character): most ATPs are run on single problems or batches. Urban et al.'s MaLARea [33] that alternates between theorem proving runs and premise selection is an exception. Gauthier et al.'s Alien Coding [7] has been generating programs that cover new integer sequences in the OEIS [26] for over a year without plateauing, which suggests that theory exploration systems such as Hipster [16] could prove beneficial if run autonomously over interactive theorem proving (ITP) systems.

Autonomy suggests another feature: maintaining a *worldview*. An AGI, even a non-autonomous AGI scientist, should maintain some model of the world with which it deals. This suggests that the AGITP[1] should live on the level of ITP systems — perhaps transferring learning among formal math libraries. Technically, ENIGMA's GNN is probably developing some sort of worldview over the whole library's solved problems. Premise selection can be seen as loading long-term *declarative memory* into *working memory*, which in (sledge)hammer [4] settings involves translating from one logical language to another. Semantic guidance of theorem provers via finite interpretations of theories [25] can be integrated into modern systems and perhaps used to flesh out comprehensive worldviews of ITP libraries. The core requirements for *conjecturing* are theory exploration and quality recognition. An AGITP system that is continually exploring the ITP libraries, refining the proofs, seeking useful lemmas, and developing a math-worldview will probably be a good foundation from which to learn how to recognize novel, interesting conjectures and lemmas.

Two crucial features for an AGI system are *self-organization and metalearning*[2] [5, 28]: metalearning is the process by which a learning algorithm learns how to learn better, such as applying one learning algorithm to fine-tune another. Metalearning goes well with self-organization and reflection where the AI, ATP, and ITP components of the AGITP system should be integrated together without the need for a human-in-the-loop to choose when and how to link them up. Set up AI components to choose when to stop an iteration of training loops, when to switch datasets, when to tweak the featurization for AI models, when to explore new strategies and parameterizations [13], etc. The option of online learning also seems important, such as with Tactician in Coq [2,3,40]. My own work with ProofWatch [9,10], Parental Guidance and 3-phase ENIGMA [11,12] suggests that finding ways to integrate additional information into the theorem proving loop, plugging AI into more *choice points* within the ATP, can significantly increase performance. The hypothesis is that ATP and ITP system performance and generality will increase as the components are effectively modularized and integrated [3].

The final capacity discussed in the abstract is the importance for an AGI to interact with the "real world" and new domains. *Autoformalization* [17–19, 31, 36, 37, 41] can help map many domains described in natural language into formal problems amenable to theorem proving. The capacity to produce formal descriptions of multi-media scenes is also important; conversely, working with geometric models could help with guidance on geometry problems or conjecturing.

[1] *AGITP*: the AGI theorem prover.

[2] Incidentally, these are features on which Large Language Model-based systems are currently weak, too.

[3] I recognize that this is difficult and that initial attempts at generalization can fail to outperform human researchers. To this end, I suggest setting up a general ecosystem in which AI systems can compete with humans on each subtask, so that the transition to full autonomy will be smooth as AI techniques rise to the challenge.

Acknowledgements. This work is supported by Dar CISCO No. 2023-322029 A Formalization of Ethics for Decision Support.

References

[1] Alexander Bentkamp, Jasmin Blanchette, Sophie Tourret, and Petar Vukmirović. Superposition for full higher-order logic. In André Platzer and Geoff Sutcliffe, editors, *Automated Deduction – CADE 28*, pages 396–412, Cham, 2021. Springer International Publishing.

[2] Lasse Blaauwbroek. The tactician's web of large-scale formal knowledge, 2024.

[3] Lasse Blaauwbroek, Josef Urban, and Herman Geuvers. The tactician. In Christoph Benzmüller and Bruce Miller, editors, *Intelligent Computer Mathematics*, pages 271–277, Cham, 2020. Springer International Publishing.

[4] Sascha Böhme and Tobias Nipkow. Sledgehammer: Judgement Day. In Jürgen Giesl and Reiner Hähnle, editors, *IJCAR*, volume 6173 of *LNCS*, pages 107–121. Springer, 2010.

[5] Tyler Cody and Peter A. Beling. Towards a process algebra and operator theory for learning system objects. In Kristinn R. Thórisson, Peter Isaev, and Arash Sheikhlar, editors, *Artificial General Intelligence*, pages 43–52, Cham, 2024. Springer Nature Switzerland.

[6] M. Gagliolo. Universal search. *Scholarpedia*, 2(11):2575, 2007. revision #152144.

[7] Thibault Gauthier, Miroslav Olšák, and Josef Urban. Alien coding. *International Journal of Approximate Reasoning*, 162:109009, 2023.

[8] Ben Goertzel, Vitaly Bogdanov, Michael Duncan, Deborah Duong, Zarathustra Goertzel, Jan Horlings, Matthew Ikle', Lucius Greg Meredith, Alexey Potapov, Andre' Luiz de Senna, Hedra Seid Andres Suarez, Adam Vandervorst, and Robert Werko. Opencog hyperon: A framework for agi at the human level and beyond, 2023.

[9] Zarathustra Goertzel, Jan Jakubův, Stephan Schulz, and Josef Urban. ProofWatch: Watchlist guidance for large theories in E. In Jeremy Avigad and Assia Mahboubi, editors, *Interactive Theorem Proving - 9th International Conference, ITP 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 9-12, 2018, Proceedings*, volume 10895 of *Lecture Notes in Computer Science*, pages 270–288. Springer, 2018.

[10] Zarathustra Goertzel, Jan Jakubův, and Josef Urban. ENIGMAWatch: ProofWatch meets ENIGMA. In Serenella Cerrito and Andrei Popescu, editors, *Automated Reasoning with Analytic Tableaux and Related Methods - 28th International Conference, TABLEAUX 2019, London, UK, September 3-5, 2019, Proceedings*, volume 11714 of *Lecture Notes in Computer Science*, pages 374–388. Springer, 2019.

[11] Zarathustra A. Goertzel, Jan Jakubův, Cezary Kaliszyk, Miroslav Olšák, Jelle Piepenbrock, and Josef Urban. The Isabelle ENIGMA. In June Andronick and Leonardo de Moura, editors, *13th International Conference on Interactive Theorem Proving (ITP 2022)*, volume 237 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:21, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[12] Zarathustra Amadeus Goertzel, Karel Chvalovský, Jan Jakubuv, Miroslav Olsák, and Josef Urban. Fast and slow Enigmas and Parental Guidance. In Boris Konev and Giles Reger, editors, *Frontiers of Combining Systems - 13th International Symposium, FroCoS 2021, Birmingham, UK, September 8-10, 2021, Proceedings*, volume 12941 of *Lecture Notes in Computer Science*, pages 173–191. Springer, 2021.

[13] Jan Hůla, Jan Jakubův, Mikoláš Janota, and Lukáš Kubej. Targeted configuration of an smt solver. In Kevin Buzzard and Temur Kutsia, editors, *Intelligent Computer Mathematics*, pages 256–271, Cham, 2022. Springer International Publishing.

[14] Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

[15] Jan Jakubuv and Josef Urban. ENIGMA: efficient learning-based inference guiding machine. In Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke, editors, *Intelligent Computer Mathematics - 10th International Conference, CICM 2017, Edinburgh, UK, July 17-21, 2017, Proceedings*, volume 10383 of *Lecture Notes in Computer Science*, pages 292–302. Springer, 2017.

[16] Moa Johansson, Dan Rosén, Nicholas Smallbone, and Koen Claessen. Hipster: Integrating theory exploration in a proof assistant. In Stephen M. Watt, James H. Davenport, Alan P. Sexton, Petr Sojka, and Josef Urban, editors, *Intelligent Computer Mathematics*, pages 108–122, Cham, 2014. Springer International Publishing.

[17] Cezary Kaliszyk, Josef Urban, and Jiri Vyskocil. Improving statistical linguistic algorithms for parsing mathematics. In Boris Konev, Stephan Schulz, and Laurent Simon, editors, *IWIL-2015. 11th International Workshop on the Implementation of Logics*, volume 40 of *EPiC Series in Computing*, pages 27–36. EasyChair, 2016.

[18] Cezary Kaliszyk, Josef Urban, and Jiří Vyskočil. Automating formalization by statistical and semantic parsing of mathematics. In Mauricio Ayala-Rincón and César A. Muñoz, editors, *Interactive Theorem Proving*, pages 12–27, Cham, 2017. Springer International Publishing.

[19] Cezary Kaliszyk, Josef Urban, Jiří Vyskočil, and Herman Geuvers. Developing corpus-based translation methods between informal and formal mathematics: Project description. In Stephen M. Watt, James H. Davenport, Alan P. Sexton, Petr Sojka, and Josef Urban, editors, *Intelligent Computer Mathematics*, pages 435–439, Cham, 2014. Springer International Publishing.

[20] J. Laird. *The Soar Cognitive Architecture*. Mit Press. MIT Press, 2012.

[21] John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4):13–26, Dec. 2017.

[22] Christian Lebiere, Peter Pirolli, Robert Thomson, Jaehyon Paik, Matthew Rutledge-Taylor, James Staszewski, and John R Anderson. A functional model of sensemaking in a neurocognitive architecture. *Comput. Intell. Neurosci.*, 2013:921695, November 2013.

[23] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, Dec 2007.

[24] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1244–1259, Paris, France, 03–06 Jul 2015. PMLR.

[25] William McCune. Semantic guidance for saturation provers. In Jacques Calmet, Tetsuo Ida, and Dongming Wang, editors, *Artificial Intelligence and Symbolic Computation*, pages 18–24, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[26] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences. Published electronically at http://oeis.org.

[27] James T. Oswald, Thomas M. Ferguson, and Selmer Bringsjord. A universal intelligence measure for arithmetical uncomputable environments. In Kristinn R. Thórisson, Peter Isaev, and Arash Sheikhlar, editors, *Artificial General Intelligence*, pages 134–144, Cham, 2024. Springer Nature Switzerland.

[28] T. Schaul and J. Schmidhuber. Metalearning. *Scholarpedia*, 5(6):4650, 2010. revision #91489.

[29] Artem A. Sukhobokov, Evgeny Belousov, Danila R. Gromozdov, Anna S. Zenger, and Ilya A. Popov. A universal knowledge model and cognitive architecture for prototyping AGI. *CoRR*, abs/2401.06256, 2024.

[30] G. Sutcliffe. The TPTP Problem Library and Associated Infrastructure. From CNF to TH0, TPTP v6.4.0. *Journal of Automated Reasoning*, 59(4):483–502, 2017.

[31] Christian Szegedy. A promising path towards autoformalization and general artificial intelligence. In Christoph Benzmüller and Bruce Miller, editors, *Intelligent Computer Mathematics*, pages 3–20,

Cham, 2020. Springer International Publishing.

[32] Kristinn R. Thórisson and Gregorio Talevi. A theory of foundational meaning generation in autonomous systems, natural and artificial. In Kristinn R. Thórisson, Peter Isaev, and Arash Sheikhlar, editors, *Artificial General Intelligence*, pages 188–198, Cham, 2024. Springer Nature Switzerland.

[33] Josef Urban, Geoff Sutcliffe, Petr Pudlák, and Jiří Vyskočil. MaLARea SG1 - Machine Learner for Automated Reasoning with Semantic Guidance. In Alessandro Armando, Peter Baumgartner, and Gilles Dowek, editors, *IJCAR*, volume 5195 of *LNCS*, pages 441–456. Springer, 2008.

[34] Joel Veness, Kee Siong Ng, Marcus Hutter, William Uther, and David Silver. A monte-carlo aixi approximation. *J. Artif. Int. Res.*, 40(1):95–142, jan 2011.

[35] Peter Voss and Mladjan Jovanovic. Concepts is all you need: A more direct path to agi, 2023.

[36] Qingxiang Wang, Chad Brown, Cezary Kaliszyk, and Josef Urban. Exploration of neural machine translation in autoformalization of mathematics in mizar. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, CPP 2020, page 85–98, New York, NY, USA, 2020. Association for Computing Machinery.

[37] Qingxiang Wang, Cezary Kaliszyk, and Josef Urban. First experiments with neural translation of informal to formal mathematics. In Florian Rabe, William M. Farmer, Grant O. Passmore, and Abdou Youssef, editors, *Intelligent Computer Mathematics*, pages 255–270, Cham, 2018. Springer International Publishing.

[38] Christoph Weidenbach. Combining superposition, sorts and splitting. In John Alan Robinson and Andrei Voronkov, editors, *Handbook of Automated Reasoning (in 2 volumes)*, pages 1965–2013. Elsevier and MIT Press, 2001.

[39] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[40] Liao Zhang, Lasse Blaauwbroek, Bartosz Piotrowski, Prokop Černỳ, Cezary Kaliszyk, and Josef Urban. Online machine learning techniques for coq: A comparison. In Fairouz Kamareddine and Claudio Sacerdoti Coen, editors, *Intelligent Computer Mathematics*, pages 67–83, Cham, 2021. Springer International Publishing.

[41] Claus Zinn. *Understanding informal mathematical discourse.* PhD thesis, University of Erlangen-Nuremberg, 2004.