The Reasoner

Volume 17, Number 1
January 2023

thereasoner.org
ISSN 1757-0522

Contents

Editorial

Dear Reasoners, it a pleasure to introduce my interview with Josef Urban. Josef is distinguished researcher at the Czech Institute of Informatics, Robotics and Cybernetics (CIIRC), and an expert in combining automated theorem proving with machine learning. In the interview, we touch on a variety of issues connecting mathematics, humans, and machines – including the revival of a 1999 speculation by John McCarthy concerning the extension of proof-checking methods to arguments-checking in the public arena, as well as open bets on the future of AI-driven theorem proving.

At a time in which explainable AI is a pressing intersectoral challenge, it is very fascinating to hear about the role of machine learning in pushing forward our ability to mechanise mathematical theorem proving, which is a prototypical example of both human ingenuity and its ability to articulate reasoning in a transparent way.

Before I leave you to it, let me thank Josef warmly for his time and for his willingness to share many interesting details about his personal trajectory in academia, so far.

Hykel Hosni
University of Milan

Features

Interview with Josef Urban

Hykel Hosni: Can you tell us about your background?

Josef Urban: I studied math (MSc) and economics (Bc), and then computer science (PhD) in Prague. I started in 1992 which was three years after the Velvet revolution in Czechoslovakia. It was a time of political and economic changes in which we got involved in 1989 as students of a mathematically oriented high school in Prague. In 1992 I decided to study math simultaneously with economics. I wanted to learn more about math and its foundations because things like Gödel's theorems looked like deep mysteries. But I was also interested in the social sciences that related to what was happening in Czechoslovakia and similar countries. My Bc thesis in economics was analyzing the inflation in the post-revolution years - working with a lot of data and coming up with conjectures explaining them.

HH: And what about your thesis in Mathematics?

JU: I was from the start attending lectures and seminars by

Petr Vopenka, who had just returned to teaching in 1992 after being the post-revolution minister of education. I enjoyed his lectures on the foundations of mathematical thinking and his books and seminars about the history and philosophy of math and science. Quite early, as I was digesting the bottom-up development of various math theories and trained to solve harder and harder exercises, I started to wonder how much of the process of "solving math problems" could be mechanized. In my MSc I eventually specialized in logic and set theory, learning about Gödel's incompleteness, undecidability, forcing, and non-standard set theory.

HH: Did you see mathematics as a mechanical, rather than creative, activity back then?

JU: Despite learning about Gödel's, Church's, and Turing's negative results in these lectures, I was still wondering how humans become experts in math. It felt like the combination of studying the theory, previous problems and solutions, exploring on my own, and learning from the successes and mistakes, eventually made me reasonably competent.

HH: The "exploring on your own" bit doesn't sound too machine-like though.

JU: I had read Penrose's books early (and enjoyed the physics there) but never found his arguments for human supremacy in math convincing. "Exploration" has been a particularly interesting AI topic, with very simple heuristics for balancing "exploration" and "exploitation" being recently one of the reasons behind the success of AI systems playing Go. Much later I read Turing's 1950 paper *Computing Machinery and Intelligence* which systematically goes over the arguments against thinking machines and largely dismantles them. I recommend it to anybody interested in AI. Especially to those logicians who may (with Penrose) believe that Gödel's theorems and Turing's 1936 proof of undecidability somehow show that machines cannot do math on the (super)human level.

HH: Did you write your MSc dissertation on this topic?

JU: Sort of. I decided to work on automated reasoning, and found Petr Stepanek who agreed to supervise me. He had a wide range of interests, from logic and set theory, to lambda calculus, Prolog, and automated reasoning. During my MSc, I learned about standard resolution and tableaux-based theorem proving, learned to program in Prolog and Lisp, and read Art Quaife's PhD thesis on Automated Theorem Proving in set- and other theories. I also read Doug Lenat's papers on the Automated Mathematician (AM), Eurisko and Cyc, as well as Alan Bundy's book on computer modeling of mathematical reasoning.

HH: That surely formed a solid basis for postgraduate work. How influential has this been in your later work?

JU: I agreed with Bundy that domain-specific heuristics are important in problem-solving: uniform (even if theoretically complete) procedures like resolution are not enough, and lead to combinatorial explosion. I liked Lenat's AM, but his ultimate decision to manually encode all of the world's (problem-solving) knowledge and heuristics in Cyc did not look right to me.

HH: Why?

JU: My feeling was that humans are capable of learning the domain-specific intuitions and heuristics for problem-solving, and so computers should be able to do that automatically too. So I started to study machine learning in addition to machine reasoning, and my MSc thesis was my first (not very successful) attempt to bring the two approaches together, by mining some reasoning heuristics from a large math corpus.

HH: Which you kept doing for your PhD.

JU: Yes. After my MSc attempts, it was clear that a lot of work needed to be done. Petr Stepanek dryly stated that "dreams are killed by being pursued" and took me also for a PhD. It was about combining machine reasoning and learning over a large formal math corpus. I experimented with the Mizar Mathematical Library (MML) which was perhaps the largest piece of mathematics understandable to computers at that time.

HH: Can you tell us a bit more about Mizar?

JU: It is one of the first formal math projects, distinct by its early focus on human-like math language and proof style, and on building up a large textbook-like math library. It was started in 1973 by Andrzej Trybulec, a Polish mathematician (his PhD was in topology under Borsuk) with an interest in linguistics, logic, and computers. His plan in 1973 was to "quickly" develop a system that would proof-check his PhD thesis, and then return to math. You can easily infer that he was also into sci-fi – Mizar will turn 50 this year, and we are still not there.

HH: Well, some dreams do survive their pursuit! Was Mizar your only option at the time?

JU: There were other projects, but none of them focused so clearly on building a large textbook-like formal math library in a setting that I could understand easily.

HH: How about now?

JU: Today, we have several systems and libraries suitable for such research, e.g., Isabelle, HOL, Coq, Metamath, and Lean.

HH: Do you see any conceptual or methodological analogies between the Cyc and Mizar projects?

JU: I can't say much about Cyc since I have never really worked with it. I don't really know how they automate reasoning, what exactly the formalism is, etc. I know more about SUMO (the Suggested Upper Merged Ontology) which is an open-source common-sense reasoning project inspired by Cyc. One similarity between these and formal math projects is that a lot of interesting research is done to choose the exact logical formalism. It is a bit like developing programming languages - there is no perfect one and as you solve more tasks, the language, the underlying formalism, and the automation methods are changing. Andrzej made this at some point into his motto: "Experience, Not Only Doctrine" (citing Kreisel). It meant that designing a system like Mizar needs to be complemented by simultaneously building a large formal math library in it. It is in a way a very experimental science. Only after there is a sizeable body of formal math, can we experimentally measure how various linguistic, logical, and automation features make the formal statements and proofs easier to write and more natural.

Actually, one interesting high-level connection with common-sense reasoning is via John McCarthy, who interacted with Mizar and Andrzej at the second QED workshop in Warsaw in 1995. The practical feasibility of projects like Mizar and their aspirations expressed in the QED Manifesto inspired him to write a futuristic note about using common-sense formal reasoning to improve politics, lawmaking, etc. I really like the idea and think that it could help quite a bit especially in today's Twitter era. In a similar way as we dream about formal verification of all code on GitHub (especially when you decide to use some), we could dream about formal checkers of (important) tweets, blog posts, Wikipedia articles, etc. An obvious dissimilarity is that formal math is "easy" in the sense that there are well-established formal systems and we do have proof checkers verifying very long proofs in complex theories

today. Common-sense reasoning feels much fuzzier to me and it is hard to get any large and interesting corpora of long reasoning chains there.

HH: To the list of challenges here I'd also add that lots of people don't seem to be using social media to do any reasoning at all! Going back to your PhD. How did it go?

JU: I translated the whole Mizar library to a format used by state-of-the-art automated theorem provers (ATPs) and tried to prove about 30000 top-level Mizar theorems automatically. Andrzej was rather pessimistic, but it turned out that ATPs could prove about 40% of the theorems when given only the premises used in the human-written Mizar proof.

HH: I guess this was both surprising and very encouraging.

JU: It was indeed. However, automatically proving a theorem without such manual pre-selection of relevant facts would be much harder. On average this would mean thousands of theorems given to the ATPs, which were at that time not designed for that, typically working on problems with only a few axioms.

HH: And this is where you thought machine learning could chip in.

JU Exactly. For each of the 30000 theorems, I trained a simple (naive Bayes) statistical learner on all proofs of the previous library theorems. The task was to rank (recommend) the previous theorems based on their estimated relevance to the symbols in the conjecture. The results were again surprising. On average, the first 100 premises recommended by the trained system covered 70% of the premises used in the human-written Mizar proof. Given all the stories about undecidability and arbitrary hardness of math, this corpus (the best formal approximation to a collection of math textbooks) was surprisingly predictable. Maybe mathematicians are humans after all, and not particularly algorithmically incompressible ones?

HH: It reminds me of the line, usually attributed to von Neumann, to the effect that one does not understand mathematics, it just gets used to it. And if you have lots of examples to look at, maybe it isn't too hard to get used to it.

JU: Sounds also like what they say about quantum physics. But yes, mathematics has several interesting aspects. The intuitions of a few human experts can be developed quite far and in ways that nobody outside their circle easily understands.

HH: Do you have any examples in mind?

JU: I guess the recent extreme and controversial example is the abc conjecture. Or even just the foundations of math - how "understandable" is the fact that a particular natural phenomenon is ultimately modeled as some "set" or "type" in some abstract universe? Also, we have large formally verified proofs like the proof of the four-color theorem and Bob Veroff's impressive 100k-line ATP proofs of open conjectures in loop theory, which we know to be true thanks to the computers, but it seems wrong to say that we "really understand" why. Then there is the universality aspect of math. Galileo said that "Mathematics is the language of Science", suggesting that it allows us to express any "real-world" problem or theory we may encounter. This in turn suggests that having strong automated mathematicians could lead to large general progress in other sciences.

And then there is the undecidability aspect. Thanks to Gödel, Church, and Turing we also know that mathematical problems can be arbitrarily hard, impossible to decide, and that there is no terminating decision procedure for all of math. With algorithmically incompressible examples like Chaitin's. And yet, we are today training AI systems on mathematical theories that we (the AI/TP researchers) sometimes do not even properly understand and the systems are becoming quite competent. To make it very concrete (even if quite bold given where we are today), it could be the case that for the abc conjecture, we will have an AI system creating a formal proof (or at least a proof sketch) out of Mochizuki's book automatically, before the mathematical mainstream forms its final opinion about Mochizuki's proof.

HH: If you're right, people well outside number theory circles will speak of the abc conjecture! But I'm sure our readers, especially the graduate students, are now very curious to know how far you could push your ideas during your PhD.

JU: I did the first experiment with combining the learning and proving approaches. For each Mizar theorem, I let the trained learning system select the 30 most relevant premises for the ATP, which then attempted to find the proof. The success rate was about 15%, and I got many ATP proofs that used different premises than the human-written proofs. Some of them were also shorter. To see where this all went, fast forward to 2021. In our last evaluation (https://github.com/ai4reason/ATP_Proofs) over a much bigger recent version of MML (60000 theorems), we could prove 75% of the theorems from their manually selected premises, and 60% of the theorems when the ML/ATP system selects the premises itself - typically by some form of learning. The ATPs (and ITPs) have also recently integrated various learning methods for guiding their core inference algorithms, which was crucial to get these results.

HH: That sounds impressive. What were your moves after defending your PhD?

JU: In 2006 I got a Marie-Curie postdoc to work with Geoff Sutcliffe on further developing this new large-theory reasoning topic and combinations of reasoning with machine learning. Geoff has been a long-time developer of the TPTP library - a benchmark with thousands of problems for benchmarking automated theorem provers (ATPs). And I could suddenly easily create tens of thousands of interesting ATP problems from Mizar in the most straightforward way, and millions of additional harder/easier problems by chasing the large Mizar inference graph. This became known as the MPTP - Mizar Problems for Theorem Proving. Because it was too big for the ATPs at that time, the first realistic AI/TP benchmark we created - the MPTP Challenge - was a selection of 252 Mizar theorems and another 1000 necessary formulas (definitions, etc.) leading to the proof of the Bolzano-Weierstrass theorem. We have benchmarked the existing ATPs on it, and I started to create AI/TP systems that could solve as many MPTP Challenge problems as possible in a given global time limit (e.g. one day).

HH: Could you illustrate the basic idea behind one of those systems?

JU: Sure. The first such system became known as the Machine Learner for Automated Reasoning (MaLARea). It sounds awfully generic today, which I guess shows that this was not a particularly crowded research field back in 2006. Its idea is a pretty simple positive feedback loop between learning and reasoning: A set of problems is attacked initially with some base reasoning system, which can solve some fraction of the problems, typically the easier ones. Then a machine learning system is trained on the solutions, possibly also on the failures, and used in some way to guide the reasoning system, which is again run on the problems in the next iteration of the loop. Thanks to the learned guidance, the reasoner can now typically solve some more problems and possibly in new ways, and we

can again train the machine learner on the solutions. Etc. - this loop goes on *ad infinitum*. In the first MaLARea, this loop was instantiated by using the best first-order ATPs as the reasoning systems, and by learning which premises from the many available are the ones that are necessary for the proofs and should be given to the ATPs. Since today's ATPs are typically trying many different things – sometimes too many –, there is also the aspect of exploring and learning from successful explorations. This simple loop, already in this first setting, has turned out to work pretty well, outperforming standard ATPs, which were MaLARea's components, on such large-theory benchmarks. And a lot of what we are doing today still instantiates this basic loop in various ways.

HH: And then you left Miami to go back to Prague.

JU: Yes, after I finished the Marie-Curie project, I returned to my assistant prof. position in Prague. But the chair of the department changed, we had different opinions and several conflicts, and I and my main collaborator Jiri Vyskocil left. My general feeling is that the Czech academia has progressed since, but there are still many issues here, as in other post-communist countries, compared to countries like the Netherlands.

HH: Which is where you moved after you left Prague again.

JU: I got a postdoc position at Herman Geuver's group at RU Nijmegen, where I stayed from 2009 to 2015. Herman's group included also people like Freek Wiedijk and Henk Barendregt - all quite involved in formal math. My first job was to put formal math on the web in the form of wikis, which we however soon changed into distributed version control systems like Git with various – you could today say "GitHub-style" – hooks for verifying and hyperlinking the formal math articles. And I also started collaboration on AI/TP topics with the machine learning group of Tom Heskes, which for some reason (not me) was sharing the same space as Herman's group. And around 2011 I started collaborating with Cezary Kaliszyk, who made AI/TP into his main research field.

HH: What did you work on?

JU: Our AI/TP research at that time included several topics. In 2010 we made our first attempts at machine learning connection tableau provers (MaLeCoP and FEMaLeCoP with J. Vyskocil and C. Kaliszyk) where the trained machine learner also guides the low-level decisions (inferences) of the ATPs. Since the ML systems are typically quite slow, especially today's large neural networks, it is more challenging to guide all the low-level inferences of the fast ATPs instead of when we just use ML for one-time premise selection followed by unguided ATPs. Also, the slower the overall AI/TP system, the fewer iterations of my favorite feedback loop we can run, producing in turn less data to learn from. Especially with data-hungry methods like neural networks, this is an issue. So it took us quite a bit of research and several iterations of such low-level ML-guided ATP systems to show that low-level ML guidance is really viable. We only did it convincingly in 2018 for the tableau-based ATPs and in 2019 for the strongest saturation-based ATPs. And even today, in a fair resource-bounded setting (like that of the MPTP Challenge), ATPs equipped with fast-yet-strong ML methods like decision tree ensembles still quite often outrun ATPs equipped with deep learning. I guess this gave me quite a bit of feedback about the different options and tradeoffs when guiding reasoning systems by learning at various levels.

HH: Is this when you came up when the Blind Strategymaker?

JU: Indeed. In 2012 I started working on a "mid-level" AI/TP system that ended up being called the Blind Strategymaker (BliStr). BliStr evolves a set of complementary high-level guiding strategies for state-of-the-art ATPs like E prover. Such ATPs are very large pieces of code with many parameters and domain-specific languages (DSLs) that can be used to program the search strategies of the prover. The advantage over the low-level ML guidance is that the AI/ML can be used only once to develop the high-level strategy (a program written in the DSL), which then guides the ATP without any further calls to the AI/ML system.

HH: Can you tell us a bit more about BliStr?

JU: The Blind Strategymaker is (of course) again a loop. Starting with some initial population of ATP strategies and a larger set of problems, BliStr co-evolves the population of the strategies with a population of solvable training problems, with the ultimate goal of having a reasonably small set of strategies that together solve as many of the problems as possible. The initial expert strategies ("predators") are mutated and quickly evaluated on their easy prey (the easy training problems they specialize in), and if the mutations show promise (faster solutions) on such training subset, they undergo a more expensive (more time allowed) evaluation on a much wider set of problems, possibly solving some previously unsolvable ones and making them into further training targets. Just randomly mutating on the so-far-unsolved problems is typically quite inefficient, so one really needs the intuition about which training data the mutations should be grown on. So we again have an "inductive" training phase, followed (if successful) by a "hard thinking" phase, in which the newly trained strategies attempt to solve some more problems, making them into further training data. The intuition and the deductive capability co-evolve again; doing just one of them does not work so well. It is a bit like evolving a population of animals to specialize in eating/hunting particular kinds of food. As they specialize, more and more food – i.e. problems– becomes available to them.

HH: Fascinating. You are depicting a very harmonious ecosystem populated by (only apparently) competing AI approaches, and by humans as well.

JU: At this point I'll interject a remark about "slow and fast thinking". Recently, with the development of deep learning, some people have argued that the trained neural nets are responsible for what Kahneman called "fast", intuitive, thinking, and we – the AI researchers– still need to work on emulating the "slow" – more abstract, high-level, mathematical, etc.– modes of thinking. It is interesting and tempting to think of analogies with the various MaLARea and BliStr-style loops and their "learning" and "guided solving" phases. But I am not sure if such analogies quite fit. In particular, we can, and do, make both the learning and the solving phases arbitrarily long in these loops, depending on the methods and systems involved and other parameters. For example, both statistical/deep learning and symbolic program learning/synthesis/evolution (which may include processes like concept creation) can be made very expensive. And on the other hand, some "mathematical/reasoning" algorithms, e.g., arithmetic, SAT-solving, ATP indexing, (anti)unification, etc., are today implemented in extremely fast ways on computers, capable of performing hundreds of thousands of operations per second. So it seems to me that the space of interesting AI solutions and complex metasystems is today much wider because we are not constrained by what is easy or hard for the human brain and often work with components,

think, e.g., of calculators, that already have some inhuman or superhuman aspects. Then again, looking at how humans do things is obviously useful, especially in fields like math and theorem proving where humans are still capable of things that computers cannot do.

HH: I'll now resist the temptation to ask you to make even bolder forecasts about the future of AI . . .

JU: But I have *real* bets to tell you about! Around 2012-2014 with Cezary, we combined some of these AI/TP approaches and showed that we can automatically prove 40% of the Flyspeck (Formal Proof of the Kepler Conjecture) and Mizar problems without any human assistance. And with Cezary, Jasmin Blanchette and my student Daniel Kuhlwein, we have added Isabelle/HOL to the list of such systems. We could also show many examples of simpler alternative proofs discovered by the AI/TP systems. These results got some attention, and it started to look like we may have a shot at ending the long period of plateauing performance in the first-order ATP field and speed up formalization. I started to get invitations to give talks on the topic, and in 2014 at a long talk at Institut Henri Poincare, I went quite far and created three concrete AI/TP bets on how far we will get and when. And I announced that I would bet up to 10000 EUR on my predictions. Tom Hales immediately reacted by saying that he could back my side some more. Both Iand Cezary got ERC funding to develop the AI/TP methods further in 2015 and 2016.

HH: I can only wish you good luck with your bets, and ask you not to forget to let us know how it goes with them! Meanwhile, tell us a bit about your AI4REASON project.

JU: In 2015 I got the ERC Consolidator grant and decided to take it to the new research institute CIIRC that was created in Prague as a part of the Czech Technical University (CTU - the largest technical university in Czechia). I liked the people and the research environment at RU Nijmegen, and still collaborate with them, but it felt like a good opportunity to establish an ERC-level research and international group in Czechia. It was the first ERC project at CTU.

HH: A nice comeback after you felt you had to leave Prague a few years earlier.

JU: The jury is still out on whether it was a good move. The Czech academic ecosystem, the university governance, the funding system and bodies, etc., has many problems, the largest being disproportionate bureaucracy. Bringing scientists with international experience sounds nice, but not if they have little influence on changing the main issues. However, the project ran from 2015 to 2020 and had the following directions: (i) developing strong premise selection methods, (ii) developing low-level ML guidance of ATPs, (iii) developing methods for inventing good lemmas and conjectures, (iv) combining the methods into feedback loops and larger metasystems, and (v) developing first ML-assisted methods for automated formalization of math (e.g., turning Latex to Mizar). The five years were probably the most intense academic time in my life - we have produced a number of new methods, systems, datasets, benchmarks, evaluations, and over 70 papers about them. It also included building up the group and helping to develop – and occasionally defend the existence of – the institute, four years of reviewing proposals for the Czech science funding agency (GACR), getting involved in the CLAIRE organization which tries to strengthen AI in Europe, etc. We have also started the AITP conference, which has grown from 30 people in 2016 to 80 people in 2019. And I have been "evangelizing" about the topic a lot, ranging from invited talks at conferences and seminars to informally consulting for and helping start AITP groups at various places. The community has grown and today there are groups and people doing AITP research at Google, OpenAI, DeepMind, Facebook, Microsoft, and top universities like Cambridge, Harvard, Toronto, Stanford, etc.

HH: Sounds like you lived up to the promises made in the proposal!

JF The results of the project are described in detail in its final report. The highlights in learning-based internal guidance of ATPs were the rlCoP and ENIGMA systems.

In 2018 rlCoP introduced AlphaZero-style learning-guided Monte-Carlo tree search (MCTS) over the connection calculus (in particular, Jens Otten's leanCoP system). It took Cezary and me several months of focused effort to make the final push, varying the ML systems and their integration in leanCoP, the features used for learning, the MCTS parameters, removing leanCoP's pre-programmed heuristics which were interfering with the ML guidance, etc. At some point, we figured out that we can very cheaply (fast hashing) decrease the dimension of the features used for training the ML guidance with practically no penalty on the quality of the trained guidance. This made everything much more efficient and allowed us to do multiple MaLARea-style proving-learning iterations over the whole Mizar. In 5 iterations, this doubled the performance of the initial untrained rlCoP, both on the training set and on the unseen test set. The improvement over leanCoP (which outperforms unguided rlCoP thanks to pre-programmed heuristics) was 42% in abstract time (i.e., using the same limit on the number of inferences) which was a lot.

ENIGMA (Efficient learNing-based Inference Guiding MAchine) has been a parallel line of attack on the strongest saturation-based ATP systems - in this case Stephan Schulz's E prover. Long story short, after about three years of research with Jan Jakubuv, who is the main ENIGMA developer, and others, we made it work too. In 2019, using similar tricks as in rlCoP, ENIGMA improved over the best E prover's strategy by 70% after six MaLARea-style proving-learning loops over Mizar (https://doi.org/10.4230/LIPIcs.ITP.2019.34). At that point, we were so confident about the efficiency of the ML guidance that we no longer used the abstract time for the comparison with E, and instead ran all – guided and unguided – provers with the same real time limit, which was 10 seconds. Showing that the strongest ATPs can be improved so dramatically by efficient low-level ML guidance has been probably the greatest practical breakthrough of the project. I should also note that Martin Suda has recently brought such methods to Vampire - the best ATP of the recent decade or two - again achieving impressive improvements.

Many other things happened. Thibault Gauthier, supervised by Cezary, developed TacticToe - a system that learns to guide the application of tactics in HOL, again using MCTS. The first release in 2017 already beat other (hammer) methods on HOL and in 2018 TacticToe proved 66% of the HOL library when given 60 seconds CPU time for each goal. This again inspired a lot of follow-up research on guiding tactical provers.

With Cezary Kaliszyk, Jiri Vyskocil, Shawn Wang and Chad Brown we have developed the first systems and datasets for automated formalization. Initially, these were probabilistic parsers that we augmented with type checking, trained on synthetic corpora of aligned informal-formal formulas, typically

attempting to (dis)prove the resulting parsed formal statement by further AITP methods – here is a demo: http://grid01.ciirc.cvut.cz/~mptp/demo.ogv. In 2018, we started to train neural networks for autoformalization and after some research they turned out to be surprisingly good at translating informal formulas to formal on the synthetic corpora, making me quite optimistic.

As we have been playing with such tools, we have realized that they are also producing many conjectures, e.g., the alternative formal parses of an informal statement, and that they have interesting capabilities for learning symbolic tasks like rewriting. This led to another branch of our research, where we started to use neural nets, in particular, what is today called *language models*, for conjecturing (creating cuts), proposing premises (premise selection), proof steps, and even whole proofs, and for learning various symbolic tasks like arithmetic and polynomial simplification, rewriting in loop theory, estimating the provability of a formula, etc. Again, this has been followed a lot recently. Both neural autoformalization and using language models for conjecturing, proof guidance and various symbolic and synthesis tasks seem to be booming topics today.

HH: Is there a development of the many branches of this project which you think is particularly interesting?

JU: One is an "invariant" graph neural network (GNN) by Mirek Olsak https://doi.org/10.3233/FAIA200244. It is much more efficient than the language models, capable of seeing many more symmetries, and also seeing analogies between theories with differently named concepts. It is currently the strongest system for premise selection, and in combination with other methods it has also improved a lot the low-level guidance in systems like rlCoP and ENIGMA. Jelle Piepenbrock has recently modified the GNN also for joint learning of the synthesis of elements of arbitrary Herbrand universes. This is possible exactly because of the GNN being invariant under symbol renaming, which Jelle complemented by problem-dependent term synthesis. In practice, this means that even when we get a problem with terminology that has never been used before, we can often make good predictions. This is quite unusual today. The prevailing language models typically just assume fixed language and train on "everything that is out there" (on the internet or GitHub), but often underperform in dynamic environments where new terminology is introduced often.

A recent fun project by Thibault Gauthier is learning-guided invention of programs for OEIS sequences (https://arxiv.org/abs/2202.11908). We have been running his feedback loop in various ways for several months now, inventing explanations for thousands of integer sequences. For primes, we got so far over twenty formulas. Some of them are clearly incorrect but still interesting, like Fermat pseudoprimes. So we have at this point about one million different formulas - proposed explanations - of various mathematical phenomena that fit the "experimental data" (the particular integer sequence). And some of them look quite unusual or surprising. And we can do interesting things with them, e.g., try to automatically prove that two explanations that look different are actually equivalent. And again, we can train our AITP systems on such datasets. And imagine that we start doing this for all kinds of math objects and theories, not just integer sequences ...

HH: In conclusion, can you tell us about your future plans?

JU: I guess too many when it comes to science. Art Quaife wrote in his thesis *"endless fun awaits us in the automated development and eventual enrichment of the corpus of mathematics"*. I could not agree more. I once gave a talk jokingly called "No one shall drive us from the semantic AI paradise of computer-understandable math and science", and it seems to me that the joke is turning into reality. People are getting interested in AI/TP for physics, e.g., thinking about autoformalizing whole textbooks, proposing physical laws, etc. The same with autoformalization and common-sense reasoning in areas like law. And I am almost convinced at this point that we will have autoformalization of large parts of math within a decade or two.

This alone will change a lot how we do math and hard sciences. I am very curious what happens when computers start proposing useful new theories, tools, materials, experiments, etc. There is (of course) a feedback loop between what we can theoretically explain and propose, and what we can practically build and experimentally explore. Once this feedback loop gets competently run by computers, we will have automation of science. I guess only sci-fi authors venture predictions beyond that point.

And since I now have a research group of (pretty good) human scientists in an eastern European country, my plans also include more down-to-earth topics like getting funding. With CLAIRE, we are trying to convince the EU leaders to invest significantly in AI in Europe. And with several other AI researchers in Prague, we are thinking about ways to create alternatives to the state-funded academia in Czechia, and in particular to attract private funding for AI research. We live in interesting times, and maybe even that is not complete science fiction.

News

Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming, 2 December

The workshop Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE) was conceived as a joint effort of the organizers of the BRIO Workshop (*Bias, Risk and Opacity in AI*), the ME&E-LP Workshop (*Machine Ethics & Explainability – the Role of Logic Programming*), and the AWARE AI Workshop (*Ethics and AI, a two-way relationship*). This was the first edition of the workshop, and it was co-located with AIxIA 2022, the 21st International Conference of the Italian Association for Artificial Intelligence. The event was held at the University of Udine on the 2nd of December and its main aim was to create a forum to discuss ideas on the emerging ethical aspects of AI, with a focus on bias, risk, explainability, and the role of logic and logic programming.

The workshop was one day long and included one invited talk and thirteen contributed papers. The organizers identified three main themes in the event and, accordingly, regrouped the accepted papers into three large sections: Logic for AI, Conceptual Views, and Technical Approaches to XAI. Minor adjustments were made to the schedule during the workshop; here, for the sake of clarity, we provide a brief overview of the talks by following the initial program.

The invited talk was delivered by Francesca Alessandra Lisi (University of Bari). In her talk "Ethics and Gender for a Responsible Research and Innovation in AI", Lisi proposed an overview of her current research. The talk was divided into three main parts: ethics in AI, gender in AI, and feminism

and AI. In the first part, Lisi emphasized the significance of trustworthiness in AI and the importance of identifying and avoiding biases in machine learning. In the second part, she focused on the specific problem of gender biases and discussed Londa Schiebinger's research program of gendered innovation to reach gender equality in AI. In the final part of her talk, she presented an ongoing research project at the University of Bari, the aim of which is to apply NPL algorithms and logic-based AI methods to analyze gender power relations and identify inequalities between women and men in our society.

In the section *Logic for AI*, the speakers presented various logical approaches to solve different problems in AI. First, G. Primiero and F. D'Asaro tackled the problem of identifying biases in AI from a proof-theoretic perspective, while M. Antonelli addressed the problem of dealing with probabilistic computation in a logical framework. After that, two works, one by X. Liu and E. Lorini, and the other by E. Kubyshkina and M. Petrolo, developed a modal approach to explainability. The former focused on binary-input classifiers, while the latter addressed the problem of epistemic opacity.

In the section *Conceptual Views*, A. Berman, E. Breitholtz, C. Howes and J.P. Bernardy focused on the role of a specific form of counterfactuals, namely enthymematic counterfactuals, in explaining predictions made by opaque models. S. Badaloni and A. Rodà presented their own experience in teaching the course "Gender Knowledge and Ethics in Artificial Intelligence" at the School of Engineering of the University of Padova. Finally, A. Knoks and T. Raleigh surveyed the recent work around the notion of explanation in AI that builds upon the philosophical debate on explanation and models in science.

Two sessions were dedicated to *Technical Approaches to XAI*. The first one opened with a talk by A. Castelnovo, R. Crupi, N. Inverardi, D. Regoli and A. Cosentini, who investigated the impact of biases in machine learning models by generating synthetic data with different bias combinations. The second talk, given by A. Castelnovo, L. Malandri, F. Mercorio and M. Mezzanzanica, concerned techniques for bias mitigation and fairness, focusing on their reliability and effectiveness over time. The session closed with a short paper by S. Dolgikh that addressed methods for the analysis of bias and fairness in unsupervised generative models.

The second session started with a talk by D. Fossemò, F. Mignosi, L. Raggioli, M. Spezialetti and F. D'Asaro, who used Inductive Logic Programming to explain neural networks models for user preference learning and performed Principal Component Analysis to reduce the dimensionality of the dataset, thus making model approximation more scalable. After that, A. Apicella, F. Isgrò and R. Prevete addressed the dataset shift problem with a focus on EEG-based brain-computer interfaces, exploiting XAI methods to improve the performance of these systems. Finally, M. Suffian and A. Bogliolo tackled the issue of fairness and bias mitigation in counterfactual approaches to XAI.

The event was successful in fostering the interaction between international scholars and received huge participation from a varied audience of computer scientists, logicians and philosophers. We can only wish that this is the first event of a series that in the long term may become a standard venue for the discussion of ethics, explainability and logic in AI.

Mattia Petrolo
Federal University of ABC
Giacomo Zanotti
Politecnico di Milano

Call for Papers

Logic for the new AI spring: special issue of *International Journal of Approximate Reasoning*, deadline 1 March..

UG in Irving Copi's Symbolic Logic

Irving M. Copi's version of Universal Generalisation (UG) in the first two editions of his *Symbolic Logic* (SL) came under criticism in the 1960's by at least four authors in five papers. Three in *The Journal of Symbolic Logic*: Donald Kalish (1966: "Comments on a Variant Form of Natural Deduction by William Tudhill Parry" Vol.31:2, p. 286), Donald Kalish (1967: "Symbolic Logic" by Irving M. Copi. Vol.32:2, pp. 252-255), and William T. Parry (1965: "Comments On A Variant Form Of Natural Deduction". Vol. 30:2 pp. 119-122). As well as two in *Logique et Analyse* : Hugues Leblanc (1965: "Minding One's X'S And Y'S", Vol. 8:31 pp. 209-210) and John G. Slater (1966: "The Required Correction To Copi's Statement Of UG". Vol. 9:34 p.267.)

The error is exhibited in the passage from step 3 to step 4, below:

1.(E x)(y)Fxy / ∴ (x) Fxx	
*2.(y)Fxy	Assumption
*3.Fxy	2, UI
*4.(x)Fxx	3, UG
5.(y)Fxy> (x)Fxx	2-4 C.P.
6.(x)[(y)Fxy > (x)Fxx]	5, UG
7.(y)Fxy	1, EI
8.(y)Fxy > (x)Fxx	6, UI
9.(x)Fxx 7,8,	M.P.

Copi adopts Kalish's suggestion and declares the inference from 3 to 4 , or in general, a step from

aa. Fx&Gy

to

bb. (x)(Fx&Gx)

unsound.

However, the move from aa to bb is clearly sound in Copi's system unless either 'x' or 'y' is free within the scope of an assumption in which the step is taken. For in Copi's system a variable freed by EI is always a part of an assumption. The restriction that Copi has on UG disallows generalization on a variable only if that variable is the variable over which UG is performed. Being bound by UG as a result of UG being performed on a different variable is not disallowed. But it should be disallowed as well. The restriction would stop the step from 3 to 4; as it should. For by binding the variable 'x' in 4, we are generalizing from one instance.

Alex Blum
Bar-Ilan University

What's Hot in ...

Statistical Relational AI: *One-shot learning*

How does human learning relate to machine learning, and what inspiration can one take from the other? This is the guiding question of *cognitive artificial intelligence*, topic of a Royal Society Hooke meeting in September, and also a major theme of the IJCLR conferences in Windsor which immediately followed it. The IJCLR (International Joint Conferences on Learning and Reasoning) aim to bring together the various strands of symbolic machine learning and were held this year for the second time, including conferences on Inductive (Logic) Programming, Neuro-Symbolic integration and Human- Like Computing.

A striking feature of human learning as opposed to the now-mainstream paradigm of deep learning is that humans are very efficient in learning from a small number of samples. It is enough to get burnt once on a hob to treat such devices with caution in the future; making five attempts at a new video game is enough to "get the hang of it", and if we see a single instance of a pattern, say a pair alice—ECILA, we can guess that the left-hand word was probably reversed and put into uppercase to obtain the right-hand one.

A typical neural network or reinforcement learning architecture, in contrast, would need tens of thousands of samples to reach anything close to such a level of mastery. And it is easy to see why: Indeed, there are dozens of interpretations of such a small sequence of events — Maybe this particular hob was broken, maybe the right-hand word is obtained by uppercasing the last letter of the left-hand word and appending CILA, and in our video game the perfect strategy could well be pressing one particular button whenever we are 2 cm from the enemy (we just haven't tried that particular tactic yet).

One explanation is that we might generally assume there to be a pattern to what we experience, and that therefore we favour more regular and symmetric models of reality. This idea has been very fruitful for symbolic machine learning, both theoretically and in practice.

It can be shown that under the assumption that the concepts to be learnt are very specific, one can efficiently induce logic programs encoding them from very few and only positive examples, as long as one assumes a prior distribution which penalises long encodings.

If one also restricts the hypothesis space to very regular concepts, one can perform very impressive one-shot learning, inducing patterns from a single example. Take the Flash Fill feature of Microsoft Excel, for instance: Faced with a tedious task of reformatting an entire column of Excel entries, you can simply reformat a single example entry and Flash Fill will guess the transformation you intended. More often than not, it will correctly induce the intended regular expression from a single training example.

What does this have to do with statistical relational AI, you might ask? I argue that statistical relational representations are perfectly suited to one-shot learning. In fact, one-shot learning is *the standard setting* for learning them. Typically, the training data of a statistical relational task, be it parameter learning over a given set of rules or induction of an entire model, consists of a single possible world ω on a domain D, and the learner will try to set the parameters to maximise the likelihood of obtaining M among all possible worlds on D. If we are learning the structure of the representation too, then the learner will also incorporate some complexity penalty to encourage a more compact representation.

It is worth pausing here and considering the scope of the task. What we are learning is no less than a probability distribution over the entire set of possible worlds on D, and what we are given to learn from is merely a single world. Put like this, it seems unbelievable that we can learn anything useful at all.

However, the key lies in the tight symmetry restriction inherent in a relational representation language. To see how, let's return to the running example of this column, now reframed as a parameter learning task. Imagine that we are also given some training data in the form of a single community, whose friendship relations and smoking habits are known.

The probabilistic logic program *Smokers and Friends* consists of the probabilistic facts

```
A   :: befriends(X,Y).
B   :: influences(X,Y).
C   :: stress(X).
```

and the rules

```
friends(X,Y) :- befriends(X,Y).
friends(X,Y) :- befriends(Y,X).
smokes(X)  :- stress(X).
smokes(X)  :- friends(X,Y), smokes(Y), influences(Y,X).
```

The relational language that we use, with a vocabulary of 5 relations and no constants, enforces a strong symmetry on our model, which is parametrised by only 3 different probabilities, A, B and C. If we would use a propositional language instead, we would be trying to estimate one version of A (and B) for each pair of nodes in our graph. Clearly, our data would never suffice for that: As we only have a single binary example (Are Anne and Bob friends or not?) for each pair of people, we could never estimate a real-valued *probability* that Anna and Bob were friends. Using the relational language, though, the task is meaningful, as we can exploit the rich network structure of our single community to find the optimal values of A, B and C.

Although learning the entire structure including the rules and the auxliary predicates (such as "stress") is no doubt a more complex task than simply optimising the parameters, very much the same reasoning applies, and the symmetry and expressive relational language are again what saves us from having to induce a hige number of rules to describe a complex example.

So, while we often see the expressivity and inherent restrictions of statistical relational languages as a *liability* for learning, in fact this might be precisely what makes our learning work at all.

Felix Weitkämper
Computer Science, LMU Munich

Mathematical Philosophy: *Inference to the Best Explanation*

"Inference to the best explanation" (IBE) is a much-discussed form of reasoning whereby, from the claim that E would best explain a given body of data, one concludes that E is probably true. IBE is arguably widespread in empirical science. For

instance, it seems to account for our belief in unobserved entities like quarks and dark matter. Marc Lange's recent work has campaigned for renewed attention to the nature and applications of IBE. Thanks in part to what I gather was a very productive coronavirus lockdown, Lange's last few years have yielded "What inference to the best explanation is not: A response to Roche and Sober's screening-off challenge to IBE" (*Theorema*, 2020), "Against probabilistic measures of explanatory quality" (*Philosophy of Science*, 2022), "Inference to the best explanation is an important form of reasoning in mathematics" (*Mathematical Intelligencer*, 2022), and "Putting explanation back into 'inference to the best explanation'" (*Noûs*, 2022, also a *Philosopher's Annual* selection for 2020).

I want to focus on yet another piece of this oeuvre: Lange's "Inference to the best explanation as supporting the expansion of mathematicians' ontological commitments" (*Synthese*, 2022). Here Lange argues that IBE occurs in pure mathematics too, where it often licenses belief in new entities in a way similar to its empirical counterpart. For instance, the story of mathematicians' acceptance of complex numbers can be told as a story of IBE. Since the existence of complex numbers would handily explain many otherwise puzzling facts (why certain classes of power series exhibit the same convergence behavior, why certain calculations involving square roots of negative numbers give correct answers, etc.), mathematicians are justified in taking them seriously.

On Lange's view, this ontology-expanding use of IBE is compatible with a range of metaphysical stances toward mathematics. To conclude that complex numbers exist isn't necessarily to take them on board as platonistic abstract objects, but just to say that they're mathematical entities in the same sort of good standing as other numbers, whatever exactly one takes that to be. "What an IBE argument in mathematics can confirm is that the potential explainer is an explainer and is on an ontological par with what is being explained" (7).

Another benefit of countenancing IBE in mathematics, for Lange, is that it helps us respond to Hartry Field's epistemological challenge (*Realism, Mathematics and Modality*, 1989). As Field famously argues, it seems we could have reliable mathematical beliefs only if we had epistemic access to mathematical objects and their properties, and it's unclear how such access is possible. Lange thinks a two-tiered epistemology featuring IBE may help solve Field's problem.

First, at the bottom level, our reliability with respect to basic mathematical facts can be explained by natural selection—e.g., we tend to have mostly true beliefs about simple arithmetic because doing so confers a survival advantage. (Our ancestors would have done worse if they'd been unable to reason correctly about the sizes of food caches or predator groups.)

This strategy seems promising as far as it goes, but appeals to fitness in the ancestral environment would seem not to touch our beliefs about modern abstract mathematics. It's at this point that IBE enters the picture. From our evolutionarily secure starting point, Lange suggests, we can use inference to the best explanation to gradually expand our theoretical knowledge and ontological commitments—from natural numbers to real numbers to complex numbers and beyond. The conjunction of these two justificatory strategies at least opens the door to a defense of many of our mathematical beliefs against Field's challenge.

You might wonder why we should be so confident in the deliverances of IBE. Given some ontology-expanding putative explanation E which mathematicians judge to be good, what makes E especially likely to be true? Lange's answer appeals to induction on a history of past successes: over time, mathematicians have become proficient at "anticipat[ing] accurately which mathematical facts have explanations and what sorts of explanations they have" (21).

On this suggestion, though, it's unclear why early applications of IBE should have mostly gone right. For example, we still believe in all the entities Euclid posited. Why didn't he make more explanatory missteps, given his relative lack of examples to draw on? (If you think Euclid's work is early enough and elementary enough to be covered by the evolutionary story rather than IBE, then substitute the first major mathematician for whom you think this isn't true.)

One possibility is that early mathematics did in fact contain many failed IBEs, but these were mostly discarded after their flaws became apparent. The problem with this suggestion is that it's hard to think of many actual instances of such reversals: by and large, almost any piece of mathematical ontology that was once widely accepted still has a place in mathematics today. There are examples of posits which proved less interesting or useful than originally hoped (e.g. Hamilton's quaternions), of objects whose ontological status was somewhat unclear upon their introduction (e.g. Dirac's delta function, Kummer's ideal numbers), and even a few cases of entities being temporarily banished only to reappear later in rehabilitated form (e.g. Newton and Leibniz's infinitesimals), but not many clear cases of ontology-expanding IBEs gone utterly wrong.

Were the early failures so complete that our forebears consigned all traces of them to the flames? Did we mostly avoid bad IBEs by sheer luck? Or is something more complex going on? Perhaps the right story is that mathematicians rarely rely on IBE alone to decide questions of ontology—rather, they take explanatory considerations into account alongside criteria like fruitfulness, naturalness, applicability and the like, and a new entity that scores highly in all these categories is unlikely to be tossed aside later. But does IBE feature less prominently in this tale than Lange would like?

William D'Alessandro
MCMP, Munich

Events

March

HPS: Integrated History and Philosophy of Science, University of South Carolina, 16–18 March.

June

LC2023: Logic Colloquium 2023, University of Milan, 5–9 June.

Courses and Programmes

Programmes

MA in Reasoning, Analysis and Modelling: University of Milan, Italy.
APhil: MA/PhD in Analytic Philosophy, University of Barcelona.
Master Programme: MA in Pure and Applied Logic, University of Barcelona.
Doctoral Programme in Philosophy: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.
Doctoral Programme in Philosophy: Department of Philosophy, University of Milan, Italy.
LogiCS: Joint doctoral program on Logical Methods in Computer Science, TU Wien, TU Graz, and JKU Linz, Austria.
HPSM: MA in the History and Philosophy of Science and Medicine, Durham University.
LoPhiSC: Master in Logic, Philosophy of Science and Epistemology, Pantheon-Sorbonne University (Paris 1) and Paris-Sorbonne University (Paris 4).
Master Programme: in Artificial Intelligence, Radboud University Nijmegen, the Netherlands.
Master Programme: Philosophy and Economics, Institute of Philosophy, University of Bayreuth.
MA in Cognitive Science: School of Politics, International Studies and Philosophy, Queen's University Belfast.
MA in Logic and the Philosophy of Mathematics: Department of Philosophy, University of Bristol.
MA Programmes: in Philosophy of Science, University of Leeds.
MA in Logic and Philosophy of Science: Faculty of Philosophy, Philosophy of Science and Study of Religion, LMU Munich.
MA in Logic and Theory of Science: Department of Logic of the Eotvos Lorand University, Budapest, Hungary.
MA in Metaphysics, Language, and Mind: Department of Philosophy, University of Liverpool.
MA in Mind, Brain and Learning: Westminster Institute of Education, Oxford Brookes University.
MA in Philosophy of Biological and Cognitive Sciences: Department of Philosophy, University of Bristol.
MA programmes: in Philosophy of Language and Linguistics, and Philosophy of Mind and Psychology, University of Birmingham.
MRes in Methods and Practices of Philosophical Research: Northern Institute of Philosophy, University of Aberdeen.
MSc in Applied Statistics: Department of Economics, Mathematics and Statistics, Birkbeck, University of London.
MSc in Applied Statistics and Datamining: School of Mathematics and Statistics, University of St Andrews.
MSc in Artificial Intelligence: Faculty of Engineering, University of Leeds.
MSc in Cognitive & Decision Sciences: Psychology, University College London.
MSc in Cognitive Systems: Language, Learning, and Reasoning, University of Potsdam.
MSc in Cognitive Science: University of Osnabrück, Germany.
MSc in Cognitive Psychology/Neuropsychology: School of Psychology, University of Kent.
MSc in Logic: Institute for Logic, Language and Computation, University of Amsterdam.
MSc in Mind, Language & Embodied Cognition: School of Philosophy, Psychology and Language Sciences, University of Edinburgh.
MSc in Philosophy of Science, Technology and Society: University of Twente, The Netherlands.
MRes in Cognitive Science and Humanities: Language, Communication and Organization: Institute for Logic, Cognition, Language, and Information, University of the Basque Country (Donostia San Sebastián).
Open Mind: International School of Advanced Studies in Cognitive Sciences, University of Bucharest.
Research Master in Philosophy and Economics: Erasmus University Rotterdam, The Netherlands.
Doctoral Programme in Philosophy: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.
MA in Philosophy: Dept. of Philosophy, California State University Long Beach.

Jobs and Studentships

Jobs

Fixed-term Assistant Professorship: Chair in Philosophy and Decision Theory in the Munich Center for Mathematical Philosophy, Germany, deadline 27 February.
Post-doc: in logic for trustworthy AI, University of Milan, Italy , deadline 18 December.