Project[1] Number: 649043

Project Acronym: AI4REASON

Project Title: Artificial Intelligence for Large-Scale Computer-Assisted Reasoning

FINAL Periodic Scientific Report
Part B

Period covered by the report: from [null] to [null]

[1] The term 'project' used in this template equates to an 'action' in certain other Horizon 2020 documentation

Summary of the major achievements since the start of the action

1 - Explain in a clear manner the work performed during the period covered by this report (the entire lifetime of your research project for the final report) along the main objectives/activities foreseen in the Description of the action. Please connect each achievement, where appropriate, with the relevant publication/conference presentation indicated in the other sections.

Please specify the outcome in terms of:
1.1 Research and technological achievements along the main objectives/activities (in line with the Description of the action)

WP1 Premise (Fact) Selection from large corpora:

1.1. First deep learning (DL) methods - convolutional neural networks (NNs) using a binary setting (joint NIPS'16 paper "DeepMath" with Google). Gave complementary results and involved the DL community.
1.2. Fast state-of-the-art gradient-boosted decision trees (GBDTs) allowing many iterations that provide increasingly precise training information - our IJCAR'18 paper "ATPBoost". A significant improvement on a smaller dataset.
1.3. Fast logic-aware graph NNs (GNNs) using several rounds of message passing - ECAI'20 paper "Property Invariant Embedding for Automated Reasoning" by M. Olsak. In our final Mizar evaluation, the GNN improves over the old baseline (KNN) by 25.2% (1089 vs 870 problems solved). This is a large improvement.
1.4. Stateful recurrent and transformer NNs combined with logic-aware augmentation methods showing good complementarity to the binary setting - LPAR'20 paper "Stateful Premise Selection", CICM'20 paper "First Neural Conjecturing". A different paradigm motivated by stateful methods in machine translation.

WP2 Internal Proof Guidance:

2.1. Learning-based guidance of the compact leanCoP prover. First, fast naive Bayes in leanCoP, achieving 15% improvement (LPAR'15). Followed by Monte-Carlo (MCTS) guiding methods (CADE'17), leading to the rlCoP (NIPS'18), plCoP (IJCAR'20) and gnnCoP (ECAI'20) systems that implement AlphaZero-style reinforcement learning (RL). After five proving/learning loops, rlCoP improves by 42% over unmodified leanCoP on the Mizar40 corpus - the first large breakthrough in learning-based internal guidance of ATPs, inspiring many followup combinations of RL and ATP.
2.2. Learning-based clause guidance for E (the best open ATP) - the ENIGMA system. First fast linear classifiers ("ENIGMA" in CICM 2017/2018), followed by fast GBDTs, slower recursive NNs and more advanced proof characterization ("ENIGMA-NG" in CADE'19), producing improvements on smaller benchmarks. The main breakthrough came from combining GBDTs with fast feature hashing for dimensionality reduction. Six 10-second proving/learning loops on the full Mizar library (58k problems) led to 70% improvement (25k vs 15k proofs) over the unmodified E. Such performance jump is unprecedented and it is a game-changer for the ATP field. Similar results in 2020 with fast GNNs ("ENIGMA Anonymous" in IJCAR'20).
2.3. Learning-based proof search for tactical interactive theorem provers (ITPs) such as HOL4 ("TacticToe" in LPAR'17 and "Learning to Prove with Tactics" in JAR'20) and Coq ("The Tactician" and "Tactic Learning and Proving" in CICM'20 and LPAR'20). This is technically challenging due to the complexity of the ITPs, however very rewarding. TacticToe combined with E proves about 70% of the HOL4 library. Inspired a number of followup works (Google, OpenAI, etc).
2.4. Hint-related methods for solving open algebraic problems, like the AIM conjecture, using sequences of Prover9 runs. The new methods already found (10k-100k-long) proofs of several open lemmas in AIM (https://www.cs.unm.edu/~veroff/AIM_REDONE/). Transfer of the hints methods to E and ITP corpora ("ProofWatch" in ITP'18). Using hint-based proof vectors for better ENIGMA guidance ("ENIGMAWatch" in TABLEAUX'19).
2.5. In our final large evaluation on Mizar (https://github.com/ai4reason/ATP_Proofs/blob/master/README.md), the 14-portfolio results grew from the 40% pre-project, to almost 60% (58.4%). This combines methods from WP 1,2 and 4. Our strongest single strategy matches in 30 seconds the 40% performance of the full pre-project 14-method portfolio.

WP3 Lemmatization and Conjecturing:

3.1. Lemmatization improving the performance of ATPs over large formal libraries (FroCoS'15).
3.2. Statistical-symbolic concept analogies for both targeted and untargeted conjecturing over large math corpora ("Initial Experiments with Statistical Conjecturing" in CICM WiP'16). Statistical-symbolic methods for proposing contradictory sets in large corpora ("Detecting Inconsistencies" in CADE'17).
3.3. Experiments with NNs (RNNs, transformers) for symbolic rewriting tasks ("Can NNs Learn Symbolic Rewriting?" CoRR'19), followed by DL teams (Facebook).
3.4. Neural conjecturing and synthesis methods, combined also with reinforcement learning. This includes "Tree NNs in HOL4" (CICM'20), "Deep Reinforcement Learning for Synthesizing Functions in HOL" (LPAR'20), and "First Neural Conjecturing Datasets and Experiments" (CICM'20), using transformers for making conjectures and full proof synthesis, often producing grammatically correct and sometimes mathematically interesting conjectures. Followed again by DL teams (Google, OpenAI, etc.).

WP4 Self-Improving AI Systems Combining Deduction and Learning:

4.1: Feedback loops between solving more problems and inventing better specialized strategies ("BliStr: The Blind Strategymaker" in GCAI'15, "BliStrTune" in CPP'17, GCAI'17, and "Hierarchical invention" in AIComm'18).
4.2: Integration of the WP1, and WP2 methods in our MaLARea AI/TP system, which uses several self-improving loops between theorem proving and machine learning. This includes learning of premise selection, internal guidance (ENIGMA) and strategy invention (BliStrTune). MaLARea has dominated by a large margin the 2018 and 2020 CASC large-theory competitions (http://www.tptp.org/CASC), proving about twice as many theorems as its underlying E prover system.

WP5 Autoformalization (AF):

5.1. First learning-based/semantic methods for AF. First, using probabilistic context-free grammars (PCFGs) integrated with HOL/Mizar-style typechecking and using proving for filtering; later including probabilities of subtrees. This led to good results on Flyspeck and Mizar corpora (IWIL'15, SYNASC'17 and in ITP'17 - "Automating Formalization").
5.2. First neural AF methods, giving surprisingly good results on a synthetic Mizar corpus (CICM'18, CPP'20) and

triggering followup work by DL teams.

1.2 Novel and/or unconventional methodologies

The project has been pioneering in connecting two major AI fields: Automated Reasoning (and specifically Theorem Proving) and Machine Learning. This produces novel methods in Automated Reasoning, as well as novel tasks and issues in Machine Learning. On a daily basis, we have been processing large formal mathematical corpora and some of the datasets we have created (the AIM benchmark) include research-level mathematics.

The main focus of the project has been to combine the learning and reasoning methods in useful and nontrivial ways, so that the learned systems help the reasoning and formalization processes. We have been using and modifying for the reasoning systems and corpora a wide spectrum of learning methods, from fast naive Bayes, k-NN and linear classifiers to gradient boosted trees, deep neural networks, probabilistic grammars and neural translation toolkits. The major issues are how to set up the learning, what features to use, how to make the learning systems cope with many alternative proofs to learn from, how to make the learning reasonably fast and adequate for guiding the theorem provers, etc.

We have worked with and modified a wide range of reasoning systems, from simpler connection-based automated theorem provers such as leanCoP to state-of-the-art saturation systems such as E, Vampire and Prover9, and also interactive proof assistants such as HOL, Mizar, Coq and Isabelle. The major novelty is equipping these systems with useful learning and guiding components, which can typically be done on various levels and with various granularity.

A complete novelty is our autoformalization research, where we had to first produce interesting aligned corpora of informal/formal mathematics, and then develop methods that combine statistical translation with semantic methods such as type-checking and theorem proving.

Finally, we have been building various novel high-level feedback loops between the learning and reasoning components, e.g. for strategy invention, large-theory reasoning and knowledge selection, and reinforcement learning of low-level proof guidance, conjecturing and proof synthesis.

1.3 Inter and cross disciplinary developments

Our research is mostly limited to the field of Computer Science, where we connect two major AI subfields. See the previous section for the novelty achieved there. Automated reasoning is however applied in many areas, from research mathematics to software and hardware verification. Some of our methods have helped to produce new mathematical results that the experts in the field consider interesting. We are producing automation methods that help to speed up all sorts of formal verification in the major formal proof assistants.

For an overall vision of strong AI combining learning and reasoning and assisting mathematics and science see my AGI'18 keynote (https://youtu.be/Zt2HSTuGBn8), ARW'13 invited paper (https://people.ciirc.cvut.cz/~urbanjo3/arw2013_submission_21.pdf), or 2021 Harvard CMSA presentation (https://youtu.be/UnYrWuOzOlc).

1.4 Knowledge and technology transfer

The project has focused mainly on purely mathematical libraries, as we believe that they provide the best corpora for developing our AI methods. That means that this is currently mostly pure research, even though related applications in formal verification of software and hardware may gradually appear. Still, the project has so far gained interest for mathematics-oriented companies such as Wolfram, and also from large companies that heavily invest in AI, such as Google. We are open to collaboration, however the most important and most exciting tasks are still clearly in pure research rather than in commercialization.

That said, there are several directions for technology transfer. Today's proof assistants such as HOL, Isabelle and Coq are increasingly used for industrial software and hardware verification, and verification of many other important designs (financial regulations, auctions, legal frameworks, etc.). Examples include the CakeML compiler verified in HOL, the CompCert compiler verified in Coq, and the seL4 microkernel verified in Isabelle. Practically all our methods are directly applicable in these ambitious projects, significantly lowering the currently high human cost of producing bug-free software. Systems such as TacticToe and Tactician developed for HOL and Coq make a lot of emphasis on their immediate usability by formalizers and proof engineers.

A recent spin-off is our work on the Proofgold (https://proofgold.org/) blockchain, targeting distributed formalization and democratization of access to formal proof.

1.5 Enhancing the immediate research environment

In 2016, we have co-established the conference on Artificial Intelligence and Theorem Proving (AITP - aitp-conference.org), which has grown to about 80 participants in 2019. This includes mathematicians, researchers in interactive and automated theorem proving, machine learning experts, computational linguistics, and general AI researchers.

Organizing the research community has greatly helped to develop the field and also our project. Along with our long-term collaborators such as Dr. C. Kaliszyk (ERC project SMART), Prof. Stephan Schulz, Prof. Geoff Sutcliffe, Prof. H. Geuvers, Prof. T. Heskes, and others, our group has hosted over 40 research visits and made many outgoing research visits. These were often researchers who became newly interested in AITP topics. I have helped to organize or consulted informally for AITP groups at Google, OpenAI, Facebook and several universities, and sent a student for an AITP research internship at Amazon.

The general topic of combining learning and reasoning has by 2021 made it into the AI mainstream and is being targeted by major EU AI networks such as TAILOR (https://tailor-network.eu/). Large AI and machine learning conferences (NIPS, ICLR, IJCAI, ICML) have started to include theorem proving topics. Machine learning methods have also become much more accepted at automated reasoning conferences (IJCAR, CADE, ITP, LPAR).

I have given about 20 invited talks about our topics since 2015, at venues ranging from main international conferences (AGI, SAT, TYPES) to workshops, Dagstuhl and other seminars, summer schools, etc. See section 1.1. for many of our methods being further taken up by other teams often employed by large AI companies, in particular when a deep-learning setting is involved. A number of other research groups have started to experiment with combining machine learning and theorem proving, and to look at topics such as autoformalization, learning-guided symbolic rewriting, conjecturing and related synthesis tasks, etc.

Our group has also played a significant role in establishing ERC-level research at CTU and CIIRC, leading by example in several aspects (see 1.6 and 1.7). Several top international researchers in AI (T. Mikolov, R. Babuska, J. Sivic, T. Sattler) unrelated to our group have been joining CIIRC since 2017, attracting further young talent.

1.6 Establishment and/or consolidation of the research group and team composition

The project and its support by CIIRC have led to the establishment of a world-class research group in AI and reasoning in the Czech Republic. About a half of the group are foreigners and we have also Czech nationals returning from western academia. English is the group's default language. This has been uncommon in Czech academia.

Highlights include hiring early in 2015 and 2016 Dr. Chad Brown - a top expert in higher-order automated theorem proving, and Dr. Jan Jakubuv, who eventually developed the ENIGMA system, responsible for our largest breakthroughs. A great boost has come from the expertise of prof. Robert Veroff (U. of New Mexico - since 2018 also affiliate research professor at CIIRC), who has been visiting us for two months every year. In 2017, good PhD students from US, Poland, Japan and the Netherlands were hired, some of them already on follow-up large national funding and jointly supervised with our colleagues in Austria and the Netherlands. In 2018, Dr. Thibault Gauthier and Dr. Martin Suda joined the project - both of them very talented young researchers coming to Prague from Austria. Dr. Mikolas Janota - a leading expert in QBF, SMT and SAT working previously at Microsoft Cambridge and the University of Lisbon - joined part-time in 2018 and full time in 2020, bringing his own ERC-CZ grant worth 1.5M EUR and growing the group further.

The group has obtained further funding helping its consolidation. This includes a part in a large national (EU-supported) Excellent Teams grant in 2017 (PI J. Urban, ~5M EUR), Google Faculty Research Award in 2017 (PI J. Urban, $35k), 2019 ERC-CZ grant (PI M. Janota, ~1.5M EUR), 2019 Junior GACR grant (PI M. Suda, ~260k EUR), and smaller involvement in several EU and national projects.

Team composition:
PI: Josef Urban (CZ)
Main Researchers (employed for at least one year): Chad Brown (US), Jan Jakubv (CZ), Thibault Gauthier (FR), Martin Suda (CZ), Jií Vyskoil (CZ)
PhD students: Zarathustra Goertzel (US), Bartosz Piotrowski (PL), Qingxiang Wang (CN), Filip Bártek (CZ)
Long-term visitors and smaller employments: Robert Veroff (US), Grzegorz Bancerek (PL), Michael Faerber (AT), Karel Chvalovský (CZ), Miroslav Olšák (CZ)
Project administration and technical support: Hana Krautwurmova (CZ)

Group members and visitors paid from follow-up projects:
Researchers: Mikolas Janota (CZ), Jan Hla (CZ), Karel Chvalovský (CZ)
PhD students: Lasse Blaauwbroek (NL), Yutaka Nagashima (JP), Jelle Piepenbrock (NL), Liao Zhang (CN)
Long-term visitors: Yuli Daune Funato (FR), John Hester (US), Zsolt Zombori (HU)

1.7 Others

The group has already played a significant role in linking CR with EU's AI communities and efforts such as CLAIRE, which works since 2018 towards making the EU into an AI superpower. I have proposed the creation of the CLAIRE office at CIIRC in 2018 and served since 2019 on CLAIRE's Extended Core Team. The CLAIRE Office at CIIRC has played since 2020 an important role in the VISION EU project that organizes the AI community in EU. I believe this is the way how to bring international levels of competence into Czech academia, cross-link CR with top AI research done in the EU, and to prevent brain-drain in eastern EU countries like CR. Fully aware of the major role of ERC in all this and in my personal decision to return to CR, I have also organized a joint letter in support of ERC and its budget by the three main European AI organizations (EurAI, CLAIRE, ELLIS) in 2020. I also believe that ERC and its world-class reviewing could become even more important in the process of consolidating the European AI scene.

2 - Indicate what you would consider to be the three most significant achievements in your project (e.g. the three most important scientific papers, awards, prizes, patents, interactions with stakeholders such as industry or policy makers, media reports or events, etc.).

.1. The clearest breakthrough is the ENIGMA system developed mainly by J. Jakubuv, achieving in 2019 a 70% improvement over the state-of-the-art E prover in a single-strategy setting on the large Mizar mathematical corpus. This is done by equipping E with efficient learning-based guidance trained iteratively on a growing body of proofs. The best non-learning theorem proving strategy of E can within a 10 second time limit prove 14933 of the Mizar library theorems. After six rounds of learning and obtaining more proofs, the final learning-guided strategy proves within a 10 second time limit 25397 of the Mizar theorems. This 70% increase is a very high improvement in the theorem proving domain. In the paper ("Hammering Mizar by Learning Clause Guidance", ITP 2019) we assess the impact as follows:

"We believe that this is a breakthrough that will quickly lead to ubiquitous deployment of ATPs equipped with learning-based internal guidance in large-theory theorem proving and in hammer-style ITP assistance."

This has been since followed by further systems and experiments. In a more involved final evaluation of ENIGMA and related methods we can prove almost 60% (58.4%) of the Mizar library theorems without any human help and almost three quarters (73.5%) of them when helping the ATPs by providing human-selected facts. These are very significant improvements in the automation of formal proof over the state before the project started. Compared to our 2015 pre-project evaluation, the full proof automation became almost 50% more powerful when counting the number of theorems provable in the same time by the strongest portfolio of ATP methods. The average length of the human proofs of the additional ATP-provable theorems is 21.85 lines, compared to 10.42 lines in the 2015 pre-project evaluation. See https://github.com/ai4reason/ATP_Proofs for the details and for over a hundred nontrivial mathematical proofs found automatically that we show in more depth. Two examples are discussed in my January 2021 presentation (https://youtu.be/UnYrWuOzOlc?t=1576) for the Harvard CMSA New Technologies in Mathematics Seminar.

2. A number of related significant results and new methods introduced. This includes:
(a) the first Monte-Carlo and Reinforcement Learning theorem connection provers. Our NeurIPS'18 "Reinforcement Learning of Theorem Proving" (main co-author C. Kaliszyk) work has triggered a number of follow-up systems exploring RL for theorem proving.
(b) The TacticToe (main author T. Gauthier - "TacticToe: Learning to Reason with HOL4 Tactics" in LPAR'17) has introduced a new paradigm of learning-guided tactical search, triggering follow-up systems for HOL Light (Google) and Coq (several groups including ours).
(c) Our pioneering autoformalization work combining statistical and semantic approaches led to an invited talk by C. Kaliszyk "Automating Formalization by Statistical and Semantic Parsing of Mathematics" at ITP'17. This has wide implications, opening the prospect of large-scale deep computer understanding and assistance of mathematical and scientific writings in not so distant future. This was followed by our first neural autoformalization systems and first neural rewriting and conjecturing systems and quickly taken up by other groups at Google, Facebook, OpenAI, etc.

3. The Conference on Artificial Intelligence and Theorem Proving (AITP - aitp-conference.org) which we have started in 2016 with dr. Cezary Kaliszyk. It has grown from about 35 participants in 2016 to about 80 in 2019, regularly attracting top researchers in ITP/ATP/ML and related fields. As of 2020, the AITP topic has been adopted by AI researchers at Google, Microsoft, OpenAI, Facebook and at several major universities (Toronto, Cambridge, Harvard, etc). A number of other events and conferences have started to include papers and talks with AITP topics. I have given a number of invited talks on AITP.

Overall assessment of the achievements and success of the project

Please give 1-3 sentence answers to the following:

How did the original objectives/expectations of the research project correspond to the actual outcomes? Please explain.

The actual outcomes correspond quite well to the original objectives and expectations. What we were hoping to achieve sounded very ambitious in 2014, and getting the largest breakthrough in internal guidance was nontrivial and took almost four years. The initial intuition was however correct and there is no doubt today that combinations of learning and deduction are very powerful in theorem proving.

Please provide a summary of main results of your research project.

See Sections 1.1. and 2 for details. The main measurable breakthroughs are large increases (40-70% measured in the number of solved problems) in the power of automated theorem proving over large formal corpora. A number of new methods combining learning and proving have been developed, including direct guidance of tactical ITPs, several new approaches to premise selection and internal guidance of ATPs, conjecturing, and autoformalization.

To what extent did the research project advance the field beyond the state of the art? Would you consider it a breakthrough? Please explain why.

The project has significantly advanced the field beyond the state of the art, making several breakthroughs. Integration of learning and proving and their combinations in feedback loops have led to unusually high improvements in the performance of theorem provers over large formal corpora. The project has also opened and developed first corpora and showed very encouraging results in completely new research topics, such as autoformalization.

Are there any other impacts of the project, including to society?

The main impact of the project is in AI and reasoning. The large increase in the power of the methods will likely have an immediate impact on industrial verification projects, increasing the production of fully verified software, hardware, and other technological designs. This has in turn impact on preventing costly failures of advanced technologies on which modern society critically depends. A long-term impact of this research field is increased automation of mathematics and science.

Please indicate how the research project contributed to the careers of your team members and yourself?

I have become a Principal and Distinguished Researcher - the highest research rank at CIIRC. Since the start of the project, I had several offers to work at AI companies and other universities. Some PhD students and postdocs have already been offered internships and interesting research positions around the world. Two more senior researchers in the group have recently become part of a new selective tenure-track program at CIIRC and obtained their own research funding.

Please describe the future prospects of the research group supported by the project.

The group is very well positioned in the AITP field, and there is enough project funding. A constant worry in the Czech academia is however lack of stable institutional financing for high-quality research. Even I (the first ERC holder at CTU) do not have tenure, because there is minimal institutional financing. While we have done our part and brought excellent research to CR, the CR now needs to do its part and provide stable financing for such research. Otherwise many good researchers will eventually leave again for higher salaries and job security at western universities and AI companies.

Please indicate how well you and the research project have been supported by your Host Institution?

There is a large difference between our institute (CIIRC) and the university (CTU). Support at CIIRC has been really good and striving for a great degree of competence. It is clear that CIIRC really cares about attracting good researchers, supporting foreign students, feedback from the researchers, minimizing bureaucracy, etc. This is quite different for CTU and especially its current leader, the latter showing a low degree of competence and sometimes outright hostility towards CIIRC, its staff, and their successes, most likely for petty personal reasons. This is sad and disturbing. The Czech university law and in particular the selection procedures of top university managers need a significant upgrade, e.g., to the Dutch university governance system.

Please indicate the difference receiving this ERC project has made in your case.

ERC has made a great difference in my case. It allowed me to pursue my dream topic for five years with minimal bureaucracy, and to popularize and build a group and community around the topic. This ultimately allowed us to make the breakthroughs that we hoped for. The transferability of the grant allowed me to make my move back to the Czech Republic from the Netherlands, bringing the first ERC to CTU. The independence that it gave me allowed me to speak openly and directly about issues at CTU and in Czech academia. The independence and the possibility to leave was also a strong card in our young institute's survival.

Research expeditions

List of expeditions		
Period (start-end)	**Place**	**Purpose**

Awards and recognitions

List of awards and recognitions						
No.	**Award type**	**Title of award**	**Year**	**Recipient of award**	**Reason the award was made (if applicable)**	**Any further information / clarifications**
1	*Competition Victory*	*The CADE ATP System Competition - CASC*	*2016*	*Chad Brown*	*Satallax winning the THF competition - most theorems proved*	*co-developer with M. Faerber*
2	*Competition Victory*	*The CADE ATP System Competition - CASC*	*2018*	*Chad Brown*	*Satallax winning the THF competition - most theorems proved*	*co-developer with M. Faerber*
3	*Competition Victory*	*The CADE ATP System Competition - CASC*	*2017*	*Chad Brown*	*Satallax winning the THF competition - most theorems proved*	*co-developer with M. Faerber*
4	*Competition Victory*	*The CADE ATP System Competition - CASC*	*2020*	*Martin Suda*	*Vampire winning the FOF, FNT and TFA divisions of CSC*	*Martin Suda was a co-submitter of the system*
5	*Competition Victory*	*The CADE ATP System Competition - CASC*	*2018*	*Josef Urban*	*MaLARea Winning the LTB competition - most theorems proved*	*main developer; co-developers Jan Jakubuv, Cezary Kaliszyk, Stephan Schulz*
6	*Competition Victory*	*The CADE ATP System Competition - CASC*	*2019*	*Martin Suda*	*Vampire winning the FOF, FNT, TFA and EPR divisions of CSC*	*Martin Suda was a co-submitter of the system*
7	*Research Award*	*Google Faculty Research Award 2016*	*2016*	*Josef Urban*		

Dissemination of results to academic and non-academic audience that you would like to highlight

List of disseminations								
No.	Type of activities	Main Leader	Title	Date	Place	Type of audience	Size of audience	Countries addressed

Information on other important outputs which have arisen from this project (such as software, databases, or other types of outputs).

List of other outputs
https://github.com/ai4reason/ https://github.com/JUrban/ http://ai4reason.org/

Further information

-

List of free keywords

Artificial Intelligence, Automated Reasoning, Machine Learning, Automated Theorem Proving, Interactive Theorem Proving, Formal Verification, Formalization of Mathematics