Automated Reasoning for Explainable Artificial Intelligence*

Maria Paola Bonacina[1]

Dipartimento di Informatica
Università degli Studi di Verona
Strada Le Grazie 15
I-37134 Verona, Italy, EU
mariapaola.bonacina@univr.it

Abstract

Reasoning and learning have been considered fundamental features of intelligence ever since the dawn of the field of artificial intelligence, leading to the development of the research areas of *automated reasoning* and *machine learning*. This paper discusses the relationship between automated reasoning and machine learning, and more generally between automated reasoning and artificial intelligence. We suggest that the emergence of the new paradigm of XAI, that stands for *eXplainable Artificial Intelligence*, is an opportunity for rethinking these relationships, and that XAI may offer a grand challenge for future research on automated reasoning.

1 Artificial Intelligence, Automated Reasoning, and Machine Learning

Ever since the beginning of artificial intelligence, scientists concurred that the capability to *learn* and the capability to *reason* about problems and solve them are fundamental pillars of intelligent behavior. Similar to many subfields of artificial intelligence, *automated reasoning* and *machine learning* have grown into highly specialized research areas. Both have experienced the dichotomy between the ideal of imitating human intelligence, whereby the threshold of success would be the indistinguishability of human and machine, as in the Turing's test [12], and the ideal that machines can and should learn and reason in their own way. According to the latter ideal, the measure of success is independent of similarity to human behavior. Incidentally, the Turing's test has been often misunderstood as justifying negative statements on the feasibility of machine intelligence: contrarywise, Turing's article is enthusiastically positive in this regard.

While both automated reasoning and machine learning have made amazing strides, their perception in artificial intelligence, computer science, and general culture appears to be quite different. For example, in general culture and perhaps also in more specialized circles, machine learning is seen as almost a synonym of artificial intelligence, or, to put it another way, artificial intelligence has been reduced to machine learning. On the other hand, there seems to be little or even no awareness of the historic and conceptual ties between automated reasoning and artificial intelligence. Some computer scientists acknowledge and even propound these ties, while others view automated reasoning as a subfield of the theory of computation, or of symbolic computation. For example, the ACM computing classification system places automated reasoning under logic under theory of computation [3]. Certainly, automated reasoning has deep roots reaching into theory and logic, and automated reasoning is symbolic reasoning, as machines reason about symbols and by symbol crunching.

*Partially supported by grant "Ricerca di base 2015" from the Università degli Studi di Verona.

However, the connections with artificial intelligence are just as strong, and downplaying them seems to respond to non-essential motivations. The very same keyword "artificial intelligence" has undergone several cycles of enthusiasm and disappointment, and it seems plausible that downplaying the ties between automated reasoning and artificial intelligence might have descended in part from defensive attitudes, such as staying away from artificial intelligence when it was unpopular, or avoiding grand claims on machine intelligence, thinking machines, and alike, out of a sort of cautious understatement or even self-deprecation. Reading or re-reading early articles (e.g., [12, 7]) is often a good cure for such conundrums.

Today "artificial intelligence" is very popular again, and largely identified with machine learning. On the other hand, automated reasoning is perceived, and often represented by its very same experts, not as a forefront technology, not as the heart of artificial intelligence, but rather as a background technology, as an enabling technology, as something to be enveloped or even hidden inside other technologies, from computer-aided verification to natural language processing to planning. Another often heard message presents automated reasoning as a second-choice technology, such as when one hears a speaker saying: You can try machine learning first, but if your application is safety critical or if you need an accurate, rather than approximate, answer, then use automated reasoning. Are we satisfied with this kind of message? Can the present renewed popularity of artificial intelligence be an opportunity for rethinking the relationship between automated reasoning and artificial intelligence, and between automated reasoning and machine learning? We elaborate on these questions suggesting that *eXplainable artificial intelligence* (XAI) [8] could be a next grand challenge for automated reasoning.

2 Big Data

Much of the current perceived success of machine learning comes from the fact that the growth of the internet has made available unprecedented amount of data, collected by private companies and public institutions or volunteered by users in social networks. These massive amounts of data are called *big data*, and *big-data technology* refers to the algorithms, databases, information systems, and human-machine interfaces, that allow us to handle these huge data conglomerates. This phenomenon has led to the expectation that relatively traditional machine learning approaches can shine at their best and deliver unprecedented results, simply by being applied to quantities of data that could never be harnessed before. The current success of machine learning and its identification with artificial intelligence *tout court* in general culture and business culture seems to stem from this change of context rather than a change of paradigm or a major new discovery in machine learning itself. Neural networks and learning by examples are not new paradigms, but they are expected to do wonders because there is for the first time the possibility of applying them to such large quantities of training data.

Automated reasoning also is seen as a generator of big data, in the form of massive amounts of computer-generated formulae, big proofs, formalized mathematics, and formalized knowledge. Thus, automated reasoning has become a field of application of machine learning and data mining (e.g., [6, 13, 2]). This kind of research aims at enhancing automated theorem proving by applying machine learning to decide which axioms to select, when applying a prover to a very large knowledge base, or by learning from previous and similar proofs. This line of research generated a new conference series on artificial intelligence and theorem proving [9]. Another direction is to apply automated reasoning to the theory of machine learning by formalizing and checking proofs of theorems about machine learning [1]. Which other relationships between automated reasoning and machine learning could we explore? How about *applying automated reasoning to machine learning*? In the next section we discuss this challenge.

3 XAI as a Grand Challenge for Automated Reasoning

The currently widespread expectation of success of machine learning facilitates its application to increasingly difficult and delicate decisions, from advising medical doctors about therapies to advising judges on eligibility of convicts for parole. Perhaps less dramatic, but also loaded with potentially far-reaching consequences, is the application of machine learning to research publications: scientific articles are also big data, and providers of search engines and other academic services (e.g., [11]) openly envision applying machine learning to assist authors, editors, reviewers, readers, research evaluators, with advice on where to submit an article, how to select reviewers, whether to read it, as well as predictions on how many citations it will get. The enthusiasm for machine learning technology appears at times coupled with a negative attitude towards human decision makers, seen as biased, corruptible, sloppy, lazy, whereas machines are unbiased, uncorruptible, and ready to work forever.

Investigating the roots of these negative judgements on humans in cultures and religions throughout history is beyond the scope of this paper, but computer scientists should always be aware of the risk that computer technology be used in authoritarian ways. There is a risk that this message contrasting biased and sloppy humans with unbiased, unfathomable machines may generate a new wave of disillusionement with, or even hatred for, artificial intelligence. This would hardly be a positive development for computer science and computing as a whole, including automated reasoning. Perhaps more modestly, there is a risk that exceedingly high expectations are followed by excessive disappointment, opening another negative cycle in the perception of artificial intelligence. While one may muse that this is a reason to keep a distance between automated reasoning and artificial intelligence, it is not hard to see that an anti-AI backlash of this kind would be likely to affect negatively also automated reasoning anyway.

The good news is that this kind of landscape is already generating reactions in the computer science, artificial intelligence, and machine learning communities, where also machine learning experts challenge the notion that machines are unbiased (e.g., [5]). Rather, machines may amplify the biases that are implicitly contained in the human-generated data they are given for training, including, for example, bias based on gender, race, income, geography, affiliation, or topic. Thus, machine-generated advice or decisions may reflect and even strengthen existing biases. The intelligent machines we build may also act, sometimes disconcertingly, as mirrors.

Similar reflections, among others, have contributed to the quest for *eXplainable artificial intelligence* (XAI) [5, 8]. The idea is that the predictions, and the ensuing advice for decision making, provided by artificial intelligence, pretty much identified with machine learning, should be accompanied by *explanations*. Indeed, from an epistemological point of view, from the point of view of the scientific method, prediction without explanation does not make sense! Nonetheless, prediction without explanation is precisely what is offered by most of the current machine learning technology, whose methods are called for this reason *black-box approaches*.

What constitutes an *explanation* is still subject to debate (e.g., [10]). For example, explanation should be more than *transparency*, which means letting humans know how a machine reached a certain decision. Mere transparency may flood users with too much raw data, or may require advanced knowledge. An explanation should at least provide the human user with information on what could go wrong by following the machine's prediction or advice.

Explanation is not a new word in artificial intelligence and in automated reasoning: for example, *abduction* has been studied as a form of explanation. Interestingly, and quite possibly by mere coincidence, the notion of *explanation* is central to the paradigm of *conflict-driven reasoning*, where explanation means explaining by *inferences* a *conflict* between one or more formulae, or constraints, and a candidate model (see [4] for a survey). Can automated rea-

soning contribute to the development of a definition, or even a theory, of explanation? Could explanations be derived by *computational inference*? How can we bridge the gap between the statistical inferences of machine learning and the logical inferences of reasoning, applying the latter to extract, build, or speculate and test, explanations of the former? How can we bridge the gap between the apparently very different abstraction levels of explanation as in conflict-driven reasoning and explanation as in XAI? Such questions only begin to scratch the surface of the grand challenge of developing automated reasoning to achieve explainable artificial intelligence.

References

[1] Alexander Bentkamp, Jasmin Christian Blanchette, and Dietrich Klakow. A formal proof of the expressiveness of deep learning. In Mauricio Ayala Rincón and Cesar Muñoz, editors, *Proceedings of the Eighth International Conference on Interactive Theorem Proving (ITP)*, volume to appear of *Lecture Notes in Artificial Intelligence*, pages 1–17. Springer, 2017.

[2] Jasmin Christian Blanchette, David Greenaway, Cezary Kaliszyk, Daniel Kühlwein, and Josef Urban. A learning-based fact selector for Isabelle/HOL. *Journal of Automated Reasoning*, 57(3):219–244, 2016.

[3] Maria Paola Bonacina. Automated reasoning in the ACM computing classification system. Newsletter of the Association for Automated Reasoning, December 2012. Available at http://www.aarinc.org/.

[4] Maria Paola Bonacina. On conflict-driven reasoning. In Bruno Dutertre and Natarajan Shankar, editors, *Proceedings of the Sixth Workshop on Automated Formal Methods (AFM)*, May 2017. Associated with the Ninth NASA Formal Method Symposium; see http://fm.csl.sri.com/AFM17/.

[5] Nello Cristianini. Why did the chicken cross the road? In Oliver Ray, editor, *Proceedings of the Twenty-First UK Workshop on Automated Reasoning (ARW)*, April 2017. See https://www.cs.bris.ac.uk/~oray/ARW17/.

[6] Zachary Ernst and Seth Kurtenbach. Toward a procedure for data mining proofs. In Maria Paola Bonacina and Mark E. Stickel, editors, *Automated Reasoning and Mathematics: Essays in Memory of William W. McCune*, volume 7788 of *Lecture Notes in Artificial Intelligence*, pages 233–243. Springer, 2013.

[7] Edward A. Feigenbaum and J. Feldman, Eds. *Computers and Thought*. McGraw-Hill, New York, 1963.

[8] David Gunning. Explainable Artificial Intelligence (XAI), 2017. Seen on 1 May 2017 at http://www.darpa.mil/program/explainable-artificial-intelligence.

[9] Thomas C. Hales, Cezary Kaliszyk, Stephan Schulz, and Josef Urban. Proceedings of the Second Conference on Artificial Intelligence and Theorem Proving (AITP), March 2017. Seen on 5 May 2017 at http://aitp-conference.org/2017/.

[10] Alison Pease, Andrew Aberdein, and Ursula Martin. The role of explanation in mathematical research, 2017. Talk, Workshop on Computer-Aided Mathematical Proof (CAMP), Big Proof Program; seen on 20 July 2017 at http://www.newton.ac.uk/event/bprw01.

[11] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, and Kuansan Wang. An overview of Microsoft Academic Service (MAS) and applications, 2015. Seen on 2 May 2017 at http://www.microsoft.com/en-us/research/publication/an-overview-of-microsoft-academic-service-mas-and-applications-2/.

[12] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950.

[13] Josef Urban and Jiří Vyskočil. Theorem proving in large formal mathematics as an emerging AI field. In Maria Paola Bonacina and Mark E. Stickel, editors, *Automated Reasoning and Mathematics: Essays in Memory of William W. McCune*, volume 7788 of *Lecture Notes in Artificial Intelligence*, pages 244–261. Springer, 2013.