AI as Arational Intelligence?

Rudolf Marcel Füchslin and Dandolo Flumini

Abstract We introduce the notion of arational realisations of abilities. In computer science, rationality often refers to the concept of a rational agent, which is defined by the function or behavior of the agent. We focus in the way how this behavior is described and encoded. We build up on the definition of a rational process/method/argument by Günter Ropohl, which focuses on the interpretability and communicability of the realisation of said process/method. We compare arational with rational realisations of abilities and discuss key questions appearing if arationally realised methods are employed in data science. We further analyse the relevance of arational realisations of abilities for the generation of mathematical proofs, an extreme form of communicable arguments. We conclude that modern computer science could profit from a more intensive and more technical discussion with different branches of theoretical philosophy.

Rudolf Marcel Füchslin
Institute for Applied Mathematics and Physics, Zurich University of Applied Sciences, Winterthur, Switzerland, and European Centre for Living Technology, Venice, Italy.
✉ furu@zhaw.ch

Dandolo Flumini
Institute for Applied Mathematics and Physics, Zurich University of Applied Sciences, Winterthur, Switzerland.
✉ flum@zhaw.ch

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 8, No. 2, 2022

DOI: 10.5445/IR/1000150236
ISSN 2363-9881

1 Rational Agents and Rational Realisations of Abilities

One crucial task of a science of data science consists in the definition of its topic. If we deal with actual science, such definitions are always provisional and subject to ongoing refinement and extension. Often, one reads statements such as "Data science is the combination of applied statistics blended with machine learning and some domain knowledge." We claim that the science of data science is more than a bundle of methods and methodologies somehow packed together by their need for computational power. The science of data science should also look at the fundamental possibilities to reason about data.

Reasoning about what one is doing means being able to do something and using some language to communicate (at least some aspects of) how one is doing it. This article deals with a specific observation concerning this communication. Quite often, one can perform a task despite being unable to communicate a precise algorithm to perform said task. Simple examples are mechanical skills such as, e.g. riding a bike or downhill skiing. A first observation relevant to data science and machine learning is that even if a parent is not able to communicate to his/her child by language how to ride a bike, they may well be capable of explaining how to learn it. The teaching consists not of complete explanations or algorithms on how one performs the task but of instruction on learning the task.

This observation also holds for cognitive tasks, e.g. chess, with some refinements. The rules of the game can be unambiguously explained, and there exists, in principle, a straightforward algorithm for how chess can be played. For example, one can perform an exhaustive search of a fixed length and choose the best move, whereby "best" is characterised by some valuation scheme. However, that is not how even occasional players play chess. As discussed in detail by the US chess grandmaster and psychologist Reuben Fine, chess is a complex pattern-matching task that is based on implicit heuristics. One learns these heuristics by playing chess, which includes trial and error, but also a form of aesthetic experience where one realises comparably simple structures in potentially complicated, i.e. combinatorially rich, positions. These heuristics are not formalised up to now: There are sets of rules applied in chess programs, but it is not clear whether they are the same as those used by humans, for a recent discussion, see Gobet and Charnes (2018). As Fine points out, a strong player usually only performs an exhaustive search for a comparably small number of moves (around three), chooses the most promising positions and analyses those further. The underlying heuristics are not formalised and part of the player's style.

It is an essential goal of science and engineering to communicate results, which means being capable of doing something usually coincides with being able of communicating how one does it. In science, one should describe an experiment/algorithm/method such that it can be replicated by a (sufficiently trained and knowledgeable) scientist and industrial engineering requires documentation, blueprints, et cetera that fully describe machines, tools, and processes.

The coincidence of ability and communicability is highly desirable but comes not for free. This article asks whether seeking this coincidence is always necessary or even possible. Going further, requiring communicability may even impose limitations that hinder mimicking complex cognitive processes.

In what follows, we take up the approach of the philosopher of technology Günter Ropohl, who emphasises the role of communicability in rational argumentation. We refer to Ropohl (2002); a broader view on Ropohl's work can be gained from Ropohl (2009), where the operationalisation of the communicability in a general technological development process is discussed. Ropohl's approach is characterised by the use of general system theory as a means of describing complex entities in finitely many steps in an inter-subjectively verifiable way (we thank one of the reviewers for suggesting this concise formulation.) Note that Ropohl's approach, particularly in the context of innovation and technological construction, acknowledges the existence of non-communicable knowledge. He analyses in some detail how one can modularise the process of invention such that the role of interfaces between the different actors in such a process becomes transparent. Thereby, the role of the non-communicable part of what is often called "experiential knowledge" is denied or neglected but clarified and, in consequence, also used as a resource. As Ropohl points out, a systemic view on the process of innovation does not only support finding solutions (the operationalisation that he presents focuses on solutions that emerge by novel combinations of already existing parts). The systemic perspective also enables transparent valuations, ranging from technological and economic to ethical considerations.

Based on Ropohl's analysis, we introduce the notion of "arationality", where communicability requirements are relaxed to accommodate, for example, the mechanisms of how neural networks learn and operate. We distinguish between the two dichotomies rational-irrational and rational-arational. The former relates to the performance of rational agents, the later to the way how the ability of a rational agent can be described. We then argue that for data science, and particularly machine learning, a communicable description of how a task is learned is sufficient. It is desirable but not necessary to explain how it is performed. We conclude with a discussion of the production of mathematics, i.e. an artificial mathematician. A mathematical proof is an extreme form of communicable reasoning. However, if asked how proofs are found, mathematicians often refer to concepts such as intuition, which means there is no (satisfactory) formalised way to produce mathematics.

Rationality is a broadly used term. We point out that there is no general and all-encompassing definition of the term (see e.g. Oxford Bibliographies, Rysiew (2018)). As pointed out in the entry on rationality by P. Rysiew, the SEP (Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/) does contain entries on specific types of rationality but does not offer a general definition of the term. In what follows, we aim at rational communicability, i.e. descriptions of reasons, arguments, methods and instructions. We follow Günter Ropohl, who characterised rationality by a list of features (Ropohl, 2002). He regards three of these features as indispensable (translation by the authors):

- The argument, process, method, and mechanism can be described in an intersubjective way and allows discussion.
- The description must be complete and unambiguous.
- The description enables reproducibility and can be checked for logical coherence.

To avoid potentially misleading use of the term rationality, in what follows, we speak of a rational realisation of abilities, RRA for brief. A description of the RRA (there may exist multiple equivalent descriptions for an RRA) must have the properties required by Ropohl which are listed above. If a description of a realisation of abilities does not have these properties, we call it an arational description of abilities (ARA). We emphasise that the term RRA differs from the concept of a rational agent. A standard definition Russell and Norvig (2021) understands a rational agent "as one that acts as so to achieve the best outcome or, when this is uncertainty, the best-expected outcome". This definition focuses on the rationality of the function in terms of a valued outcome; Ropohl's approach (at least in the way we employ it) emphasises how a task is encoded.

An agent may be rational by the definition of Russell and Norvig without having a realisation or implementation that has a description that can be communicated in a manner that satisfies Ropohl's criteria. Such an agent is called a rational agent with arational realisation of abilities.
A seemingly obvious case of a rational agent with no RRA is artificial neural networks (ANN). Despite attempts toward explainable AI (XAI), one usually regards an ANN as a black box, see Gunning et al. (2019). One could argue that an ANN is implemented on a deterministic computer, and from a mechanistic point of view, it is easy to understand how an ANN maps input to output. However, although giving connections, nodes, weights, thresholds, etc. describes an ANN in a way that enables precise replication, we want to rule out this type of description as a qualifier for an RRA. We, therefore, impose a probably narrower, but in the context of computer and data science, usable definition of a description of an RRA as follows:

- R1: The description is *structured* in the sense that
  - R 1.1: The description presupposes a public language with a sufficiently well-defined syntax.
  - R1.2: The description organises phrases formed according to the syntactic rules of said language in a well-defined structure (e.g., a sequence or a tree-structure).
- R2: The description is *inductive* in the sense that the structure (R1) of a description is constructed by repeatedly applying specific rules to already obtained (less complex) arguments.
- R3: The syntactic and structural correctness of the description is *verifiable* by any entity sufficiently familiar with the underlying system.
- R4: The building blocks of the description carry some meaning and are *grounded* in the universe of discourse. Moreover, the structure of the description constitutes a relation that identifies assumptions/preconditions and conclusions of some sort. Additionally, we require that the structure and the involved primitive building blocks completely determine the overall meaning of the description. In some sense, we require the description of an RRA not to be more than the sum of its parts.

An ANN does not satisfy R4. This becomes apparent if one considers that, in general, there is no obvious way of adapting an ANN if the representation of the input is changed (except for a pre-processing that translates the new data format into the one the ANN has been trained with or a re-training of the network). This contrasts with algorithms for which various levels of description exist. Most algorithms can be expressed in pseudocode in which variables and input data have a grounding in the universe of discourse, but without specifying its detailed implementation. Exemplary pseudocode is not affected by this lack of information (one could argue that it is exactly this independence from the details of the implementation with simultaneous emphasis on grounding that defines pseudocode).

The above characterisation of the description of an RRA is left quite vague, and the meaning of "sufficient" in R1/R2 and "some" in R4 depends undoubtedly on the general setting and the involved actors. The language mentioned in R1.2 could be a natural language such as English, a formal language, or even a hybrid of these two (note that, e.g. a mathematical proof as it is given in a textbook or journal is usually a hybrid between some natural language and a more strict formal syntax). Using such hybrids makes sense and harvests the power of natural language even in the description of algorithms. In a setting with only a formal language and precisely defined semantics, the universe of discourse needs to be well-defined, leading to all the variants of the frame problem, the symbol-grounding and the frame-of-reference problem. Briefly, the frame problem refers to the difficulty of formal systems in dealing with changing environments, the symbol-grounding problem with the relation between objects and their representations and the frame-of-reference problem deals with the relationship between internal representation and external interpretation of outputs or actions. For a succinctly written account, see Pfeifer and Scheier (2001).

As pointed out, an agent may be qualified as rational in the sense of Russell and Norvig (2021) without having an RRA. In that case, we speak of an arational realisation of abilities (ARA). The term "arationality" must be distinguished from irrationality. An agent that does not qualify as rational may be regarded as irrational. The absence of a language–based description that follows R1-R4 of some abilities or functions is termed arational, thereby referring to the realisation of said ability and not the performance with which the ability is executed. In brief, irrational relates to the way a function is performed, arational to how this function is described.

2 Relevance for the Science of Data Science

One can argue that with respect to ANNs and machine learning in general, we rephrased a well-known and trivial fact. However, distinguishing between the quality with which some task is performed and the description of how the task is executed can be directly generalised to all sorts of agents, including humans. The definition of a rational agent given by Russell and Norvig includes data science tools (that these tools should be rational in the sense of giving sensible output is obvious). However, if we focus on the distinction between RRA and ARA, which is based on the description and not on performance, we see for data science several questions to be addressed in the future:

I *Is the fact that a tool is based on an RRA a necessary condition for accepting the tool as scientific?* Science as a social practice relies on communicability. In science, one would like to know why and how a tool works. As in the case of ANNs, a complete mechanistic description of an ARA may exist, but without grounding (requirement R4). We need to ponder the question in what circumstances this lack of grounding is acceptable. For science as a social process, the problem may in fact be less severe as it may look: The use of tools as black boxes is a necessary practice in today's science and technology (For example, scientists regularly use tools from statistics or numerical analysis without understanding them in all detail). However, in all these cases, a thorough understanding could be gained if one invests sufficient effort. And even if one does not understand a procedure in all detail, those who have implemented it are trusted to have this understanding. In other words, from the perspective of organisation or integrity of processes, the task of understanding may be subject to division of labour. The question is whether we can replace analytical understanding and communicability with other means that still guarantee sufficient verifiability.

II *Is a rational description of a learning/training strategy a possibility to cope with the issues raised in the preceding item?* For data science, this means (among other things): Is it sufficient for data science to have scientifically justified learning strategies for its tools but no detailed understanding (in the sense of R1-R4) of how they work?

III *Can ARAs support the development of RRAs?* More and more, different types of tools, partially based on machine learning, are used to produce mathematical proofs and formulate (phenomenological) physical laws. Despite considerable success in recent times, the question remains whether ARAs (combined with expert systems or other methods) are of real practical value outside some very specific use cases.

IV Assume that there exists an ARA. Is the existence of an according RRA guaranteed? This question splits into two subquestions:

a. Means having an ARA only that one has not yet understood the ability of the ARA sufficiently well? Is there always an RRA?

b. Are there ARAs for which according RRA do not exist in a practically relevant manner? Practical relevance is used here with the idea that it may well be possible to approximate an ARA by some very sophisticated RRA, which may be difficult to derive or may consist of a large set of rules. As an example of such an ARA, we use ordinary language itself. Although sufficiently educated humans can distinguish grammatically correct from incorrect sentences with high reliability, (usable) attempts to formalise human language did not result in more than approximations of the set of all correct sentences.

From this classification of types of realisations of abilities, one immediately derives classes of intelligence. Whereas we do not see a reasonable type of irrational intelligence, a rational/arational intelligence can be understood as a measure of the ability to solve problems with RRA/ARA. Arational intelligence is then a measure of the performance of a rational agent working with an ARA.

3 Towards an Artificial Mathematician

The topic of arationality in mathematics is of particular interest to us. In the following paragraph, we argue for the necessity of automating arational forms of reasoning to implement an "artificial mathematician."

The core of our argument is that in mathematics, theorems usually precede their proofs and not the other way around. Working mathematicians do not usually explore mathematical truth by enumerating theorems and proof searches. Instead, they conjecture statements that they then attempt to prove in varying degrees of formality (often depending on the sub-field of mathematics). Conjecturing happens in many forms. The range covers simple pattern and feature recognition (e.g., Goldbach conjecture) and includes visually inspired theorems from analysis, e.g., the mean value theorem or Jordan's curve theorem). Experienced mathematicians are able to conjecture successfully in very abstract areas, where the source of inspiration is hard to determine.

The example of Jordan's curve theorem is of specific interest. Even though the statement seems evident to most people, it is particularly hard to derive from first principles (for an excellent exposition, see Ross and Ross (2011)). Thus, Jordan's theorem is one of many examples highlighting how humans assess mathematical statements relatively independently of the theorem's (formal) proofs.

As most working mathematicians probably would confirm, the usual workflow to proving a conjecture consists of starting with a very informal proof idea and then progressing through several stages of proof sketches to finally arrive at a proof of desired rigor. The progression through proof stages is not usually a strictly linear process; the mathematician may find counterexamples to the original conjecture along the way, restarting the process with a refined conjecture or even abandoning the theorem because it turns out not to be "interesting" when reformulated.

Although complex and with an uncertain outcome, the conjecture-proof workflow seems to be very efficient in generating mathematical theorems. Moreover, the process is exceptionally fruitful for the mathematical sciences since filling up missing parts of incomplete proofs guides the mathematical community towards building new theories and, moreover, provides a means of assigning "interestingness" to mathematical statements.

A productive "artificial mathematician" needs to incorporate the conjecture-proof workflow or similar approaches. Note that, in contrast to the final proof, the conjecture-proof workflow is not, in our sense, rational:

1. It is unclear what languages are used (R.1).
2. The process is not inductive (R.2) because progressing in proof stages seems not to be a matter of simply applying predefined rules.
3. While the final product of the workflow, the rigorous proof, may be amenable to automatic verification, the intermediate steps to getting there are not (R.3).

Therefore, the capacity for arational decision-making and argumentation must be accessible by "artificial mathematicians" as long as they follow the conjecture-proof strategy.

Note that there are several attempts to formalise mathematics and mathematical practice; none of these attempts seem to work satisfactorily, however. For an overview of related frameworks, see Kaliszyk and Rabe (2020) and Wos and Pieper (2003). Also, note that an artificial mathematician, just like its human counterpart, is not intended to be able to prove all or any particular set of true mathematical sentences. Therefore, as envisioned by the authors, an artificial mathematician does not contradict the various incompleteness results of mathematics as presented, e.g., in Van Heijenoort (1967).

The question of which technologies are best suited to implement "arational mechanisms" on a computer is still open. While various approaches, such as evolutionary methods or Bayesian probability updates, seem appropriate, we believe that deep neural networks (DNNs) are a core component of future artificial mathematicians as they allow for some arational decision-making and conjecturing. An example for arational conjecturing is geometrical reasoning. If we can't imagine an object with certain properties, we may conjecture that such an object does not exist. A further example is the mean value theorem, which is visually plausible but requires some thought if one wants to prove it formally. In fact, the proof itself is not very difficult once one has formalised the conjecture. However, the jury is still out regarding whether today's paradigms are already sufficient and whether progress can be made by refining existing technologies and approaches.

In our view, the concept of arational decision-making is too rough and requires more differentiation. The question is: Can we subdivide the set of arational decision procedures into subclasses? The hope is that if we analyse subclasses, the community can achieve implementations subclass-by-subclass. Our strong suspicion is that a discussion with philosophers and experts in the field of the psychology of mathematics and engineering could be of significant value.

Even in writing this paper, the phenomenon we described occurs. As we stated above, we know that some of our notions are vague. We nevertheless claim to make some reasonable statements. First, the critical questions are whether our claim is valid, second, whether the vagueries can be eradicated, and, third, whether this is possible efficiently. If our claim is correct, however, we are confronted with the fact that we can communicate about a topic that we are unable to speak about rigorously. That imposes interesting questions about the role of rigour in argumentation. Rigour seems to be valuable, but many ways of problem-solving do not rely on rigour. This holds for many branches of human activity, including engineering, but even mathematical arguments (that aim at rigour) often start from a non-rigorous base.

In a broader context, one asks whether the attempt to formalise human thinking (e.g. in chess or mathematics) makes sense at all. We can build rational agents that outperform humans in various games. These agents are RRAs, because their computations always refer to the universe of discourse given by the game and are, therefore, easy to formalise. However, the question remains whether thinking and acting can always be "gamified" (encoded by a formal description of the actions and the environment).

We claim that these questions should be analysed in a broad context, even if the question sounds technical. Again referring to Ropohl's approach to rational arguments, we regard it as a rewarding challenge to rephrase technical topics in data science and AI in such a manner that they can be subject to a discourse that includes, for example, the rich tradition of phenomenology (here understood as a school of thought in modern philosophy.)

Our central question can be stated slightly differently with an emphasis on data science: Can we emulate human thinking and acting by an RRA, or is this only possible in some specific circumstances, usually described by terms such as "games" or "standard operating procedures"? Consequently, for data science and the people working in this emerging field, a central question for the future is to what extent the acts of doing and developing data science become a topic of data science itself.

References

Fine R (1967) The psychology of the chess player. Dover Publications. ISBN: 978-4-871878-15-9.

Gobet F, Charnes N (2018) Expertise in chess. In: Ericsson K.A. KA Hoffman R.R., W. WA (eds.), The Cambridge handbook of expertise and expert performance, Cambridge University Press, pp. 597–615. DOI: 10.1017/9781316480748.031.

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ (2019) "XAI Explainable Artificial Intelligence". Science Robotics 4(37):eaay7120. DOI: 10.1126/scirobotics.aay7120.

Kaliszyk C, Rabe F (2020) A Survey of Languages for Formalizing Mathematics. In: Benzmüller C, Miller B (eds.), Intelligent Computer Mathematics, Springer International Publishing, Cham, pp. 138–156. ISBN: 978-3-030535-18-6.

Pfeifer R, Scheier C (2001) Understanding intelligence. MIT press, Cambridge, Massachusetts. ISBN: 978-0-262661-25-6.

Ropohl G (2002) Rationalität und allgemeine Systemtheorie. Ein Weg synthetischer Rationalität. In: Karafyllis N, Schmidt J (eds.), Zugänge zur Rationalität der Zukunft, J. B. Metzler, Stuttgart, pp. 113–137. ISBN: 978-3-476453-07-5.

Ropohl G (2009) Allgemeine Technologie: Eine Systemtheorie der Technik, 3rd edn. Unoiverstitätsverlag Karlsruhe. ISBN: 978-3-866443-74-7.

Ross F, Ross WT (2011) The Jordan curve theorem is non-trivial. Journal of Mathematics and the Arts 5(4):213–219, Taylor & Francis. DOI: 10.1080/17513472.2011.634320.

Russell SJ, Norvig P (2021) Artificial intelligence: a modern approach, fourth edition edn. Pearson series in artificial intelligence, Pearson, Hoboken. ISBN: 978-0-134610-99-3.

Rysiew P (2018) Rationality (https://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0175.xml). DOI: 10.1093/obo/9780195396577-0175.

Van Heijenoort J (ed.) (1967) From Frege to Gödel. A Source Book in Mathematical Logic, 1879-1931. Harvard University Press, Cambridge. ISBN: 978-0-674324-49-7.

Wos L, Pieper GW (2003) Automated Reasoning and the Discovery of Missing and Elegant Proofs. Rinton Press, Princeton, New Jersey. ISBN: 978-1-589490-23-9.