Real-world Limits to Algorithmic Intelligence

Leo Pape[1] and Arthur Kok[2]

[1] IDSIA, University of Lugano, 6928, Manno-Lugano, Switzerland
[2] Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
pape@idsia.ch, a.kok1@uvt.nl

Abstract. Recent theories of universal algorithmic intelligence, combined with the view that the world can be completely specified in mathematical terms, have led to claims about intelligence in any agent, including human beings. We discuss the validity of assumptions and claims made by theories of universally optimal intelligence in relation to their application in actual robots and intelligence tests. Our argument is based on an exposition of the requirements for knowledge of the world through observations. In particular, we will argue that the world can only be known through the application of rules to observations, and that beyond these rules no knowledge can be obtained about the origin of our observations. Furthermore, we expose a contradiction in the *assumption* that it is possible to fully formalize the world, as for example is done in digital physics, which can therefore not serve as the basis for any argument or proof about algorithmic intelligence that interacts with the world.

1 Introduction

Recent theories of universal algorithmic intelligence [2, 3, 10, 11] consider optimal goal-directed computational agents that interact with the world. Combined with the view that the world can be considered the result of computation [e.g., 9, 12, 15], these theories have led to claims about intelligence in any agent [e.g., 6]. Based on highly general notions of computation that lie at the core of every formal system, the idea of algorithmic intelligence contributes to a serious computational science of intelligence that is based on solid formal proof. Theories of universally optimal intelligence that consider actual beings, such as humans, robots or animals, involve absolute claims about the nature of intelligence and the world. Such theories of intelligence, here called theories of absolute intelligence (TAIs), could potentially also be used to measure any intelligence relative to universally optimal intelligence [2, 3].

Since artificial general intelligence will not be created by abstract reasoning in formal languages, but by building a machine based on the insights achieved from our reasoning, the question arises what these absolute claims imply and what their value is for artificial intelligence. After all, strict proof (even in a probabilistic setting) is usually reserved for formal theories, not for actual machines. Is it possible to build actual machines that are, or even approximate the claim of universally optimal intelligence, is it possible to measure any intelligence relative to absolute intelligence, or are there hidden or maybe even wrong assumptions that invalidate the absolute claims? In this paper we investigate both the claims and the assumptions made by TAIs on a theoretical level.

Our approach is based on [4] (but is presented in a self-contained manner in contemporary language), which aims to identify how human reason works and what it limits are, before applying it in any argument. It is important to realize that this approach is different from the dominant, *skeptical* method of investigation, which proceeds to question without first reflecting on the assumptions already made by the skeptical questioner.

2 Universally optimal intelligence

Algorithmic theories of intelligence consider agents that interact with the world through actions and observations. The agents can be evaluated by measuring their ability to achieve a certain goal, or more formally, their ability to maximize some reward function (e.g., their score in an intelligence test). Usually, an agent does not know the reward function or the environment in advance, so it has to find the relation between its actions, observations and reward. When all components; the agent, its history of observations and actions, and the reward function are specified formally, the question which action to take can also be specified as a formal problem to which an answer can be computed based on solid formal proof.

The ability to provide proof for certain aspects of an intelligent machine can be useful to give a solid argument why a machine will function properly, for example to prove that a robot will never harm a human being, or in a probabilistic setting, that the chance it will do so is diminutive. However, recently developed theories of algorithmic intelligence [3, 10], not just provide methods to prove certain aspects of intelligent agents, but escalate into *absolute* claims about any intelligence. A proof that might originally make a simple claim, for example that an agent will always take the best action to achieve its goal, is turned into a claim about the *universally optimal* way to achieve any goal by any intelligence, including human beings.

These claims derive their absolute nature from the concept of a universal Turing machine (UTM, [13]), a theoretical computer that specifies the notion of a procedure in a formal language, such as mathematics or logic. Because the UTM *defines* the notion of a formal procedure, any operation that can ever be conceived of in a formal system can be performed by the UTM (although there are still fundamental limits of computability [1]). Defining the notion of an operation in a formal system in terms of the computations of a hypothetical computer leads to a remarkably general conclusion about our understanding of the world; since the laws of physics can be described as mathematical operations, everything in a world that can be described by these laws can be seen as the result of the computations of a UTM.

Based on such a general notion of computation, it is possible to specify the question which action an agent should take in terms of universal computation:

among all possible computations that produce an expected reward from the history of observations, actions and reward, select relations according to their probability of being the most likely. Assuming that the world is the result of computation, “most likely” can then be translated into “simplest” [3, 9], which amounts to “shortest to describe”, or “fastest to compute”. Naively, such relations could be found by trying all possible programs with increasing length for an increasing number of computation steps, and select the first one that produces the desired result [7]. An agent that bases its actions on the likelihood of the relations in its history of observations, actions and rewards, is the universally optimal agent for maximizing the reward. Such an agent would not only serve as an optimal problem solver, but could also be considered the most intelligent system that achieves any goal that can be specified as reward maximization. Moreover, if it is assumed that the objects of the agent’s computations can be fully formalized (e.g., as bits [14]), then TAIs provide a way to formally proof statements about the agent’s behavior in the real world, and about its degree of intelligence relative to other intelligences [e.g., 2, 3]. In the following, we will investigate the validity of the assumption that the world can be completely formally specified as the result of computation.

3 Conditions for knowledge of the world

3.1 Critical reflection

The search for knowledge often starts with the questioning of established dogma. Such an investigation soon leads to the realization that all claims are based on other claims, which are eventually based on assumptions with questionable validity. Any argument one tries to make, so it seems, can always be destroyed by identifying the underlying assumptions that cannot be accounted for. Moreover, even the finding that all claims are based on assumptions, must also be based on assumptions whose validity is unknown. This rather unsatisfying mode of reasoning is known as *skeptical* philosophy, because the skeptical questioner cannot account for the validity of his skeptical questions, or why his questions should even be taken seriously. But it is not the end of philosophy.

Instead, this realization is the start of a movement called *critical* philosophy [4], which investigates the methods used for reasoning before applying it in any argument. The critical approach reflects on the entire skeptical chain of reasoning, to realize that something important can be learned; there are certain assumptions we cannot positively proof, but can neither can deny or question, because their denial and questioning involves making the very same assumptions. Although such assumptions are still subjective (relative to the person that is doing the reasoning), they are also necessary, and can therefore serve as the starting point of a critical philosophy. A well-known assumption of this kind is the “I think” that accompanies every thought [4, 8].

3.2 Knowledge

Escalating everyday reasoning methods about limited observations to absolute claims is not a recent phenomenon; since long people have been fascinated by ideas about the beginning, extend and ultimate make-up of the universe (or more recently, the multiverse), the indestructible or undividable self (soul), and the idea of the omnipotent, omniscient being (God) (an in-depth discussion of this fascination is given in [4]). Books about these topics are still abundantly present in the popular science corners of bookstores (usually written by academics with some success in loosely related areas, such as astro- and particle physics or neuroscience), and this will probably remain so for some time. However, when we leave the domain of all possible observation, and engage into pure abstract reasoning, how can we be sure that our statements make any sense at all? When investigating the possibility of knowledge, where reason must be its own guide, a principle is required that distinguishes knowledge from speculation and belief, and delineates what can be known from what cannot become knowledge.

Here we place the search for knowledge in the context of a discussion about statements by the exchange of arguments. We will not use the form of a conversation, but rather study how knowledge can be obtained within the context of a discussion. Whether such a discussion is actually performed among people is irrelevant; it is not difficult to imagine a single person posing different arguments against each other. The main requirement for a discussion about the validity of statements can already be found in Plato's "Phaedo" [5], where Socrates asks:

> Socrates: At any rate you can decide whether he who has knowledge will or will not be able to render an account of his knowledge? What do you say?
>
> Simmias: Certainly, he will.

The affirmative answer might not be very surprising at first, but the excerpt actually provides the key for distinguishing knowledge from belief or speculation; somebody who claims to know something must be able to explain how he or she knows it. A statement that one cannot or does not want to defend is not knowledge, but speculation or belief, respectively.

Discussing a statement involves a comparison with what the statement does not (yet) say; otherwise there would be no need for a discussion. Hence, to discuss anything at all is to apply *limitation*. Without limitation of allowed statements, everything can be said, but nothing could be argued at all. To facilitate this limitation, a principle is required that regulates the statements that can be made (usually, based on the statements made before), by designating some statements as allowed (valid) and others as not allowed (invalid). Although many such rules can be thought of (e.g., "valid statements must contain the word 'red', invalid statements must not"), only a few facilitate a discussion in which the distinction between valid and invalid statements is maintained. Such a consistent system of rules can be discovered by reflecting on the regulative rules used during an argument, and revising them where they lead to conflicts about valid and invalid statements.

From the investigation of a discussion setting, it becomes clear that any claim on knowledge must be accounted for by following a set of regulative rules.

Without such a system of rules, there would be no way to distinguish statements from each other, determine which statements are valid or invalid, let alone obtain knowledge. However, a discussion does not have to be limited to statements that result from a finite sequence of reasoning steps, but can involve statements about the totality of all statements that result from a progressive chain of arguments, by reflecting on the method used in those arguments. Such totalities do however not result from an exhaustive deliberation of all involved objects or claims, but from an insight in the methods used to construct the discussed objects or claims. This way of reasoning is quite familiar in mathematics, for example the proof that the three angles of all triangles (in Euclidean geometry) sum up to 180 degrees does not result from considering all possible triangles, but from an insight in the method to construct triangles.

Based on this investigation of knowledge, it is now possible to make distinctions between claims that can be accounted for by reasoning (including claims about the falsity of statements), and claims that cannot make sense, because there are no methods that can ever account for what is claimed. More importantly, we have identified the method for considering claims about totalities, namely, by reflecting on the way the components of the totality are distinguished. Systems of regulative rules that can facilitate an exchange of arguments are readily available in logic and mathematics. Moreover, a formal definition of all possible procedures that could be used in a successive chain of argumentation is given in the UTM. Hence, we will use the UTM as the model for everything we can argue to know.

3.3 A world of objects

While we have identified the regulative principles of knowledge in the limited subject, it is not yet clear what the objects of such knowledge could be. It is not uncommon to consider the world as a collection of objects, whose properties and mutual relations can be discovered through scientific research. However, in the search for knowledge, the question arises how we arrive at the *concept of objects* in the first place. Our experience it not merely sensory, but also involves actively distinguishing objects. To make such distinctions, we apply rules that ascribe certain properties to limited parts of our observations. For example, starting from the distinction of regions with similar color in visual input, and relating those regions by certain rules, we can arrive at the concept of a moving object.

Here, we are not looking for an exhaustive list of properties used to distinguish objects, but try to identify the most basic principle that defines all objects. The distinction of different objects we observe and think about is based on the fundamental principle that an object must be distinguishable from what it is not. This principle preconditions any further distinctions between objects we can make, and is therefore not derived or induced from observations, but rather makes observations possible. As established before, the *methods* used to distinguish objects must adhere to regulative principles, and can hence be considered as computations of a UTM. That which allows for making a distinction between the object and what it is not, is called the object's *inner structure*. For example, when distinguishing regions of different color, the inner structure of those regions is their color, because color allowed for making the distinction between regions. Note that the issue here is not whether different colors are perceived, but how one could argue so, for example, by referring to photons with different wavelengths. For empirical objects "inner structure" always refers to an internal structure *in space*, for example, the inner structure of a musical note is the frequency at which air vibrates. To argue anything about an object's inner structure, this structure itself must also follow logical rules, and hence be composed of objects to which these rules apply. All objects can be completely specified in terms of the way they are distinguished from other objects; any further stipulations that do not address this distinction do also not contribute to the determination of the object.

Based on this definition of objects, it is now possible to consider *knowledge of objects*, as the result of the application of regulative principles that distinguish between objects. In other words, a subject needs to determine the object through the application of rules (whose form can be specified in mathematics and logic). This implies that observations do not start with objects as given, but with a limited subject that *determines* an object through the application regulative principles. Hence, when we formally describe an observed object, we have not given an account of the origin of our experience (in Kant's philosophy, this origin is referred to as thing-in-itself, which does *not* refer to an object behind the appearance of objects, but to the necessary thought that there must be a cause for our sensory experience, even though this cause cannot be known), but how we determined the object through our subjective principles. As a result, it is strictly impossible to obtain direct knowledge about the origin of sensory experience; any*thing* that can be known about observations is mediated by rules that define the observed objects. On the other hand, it is certain that all our observations can be considered the result of computation, not because the universe is written in the language of mathematics and logic, but because we use mathematics and logic to determine the objects we observe.

Multiple rules can be applied to distinguish increasingly complex objects and collections of objects. Although there are many rules that can be used to distinguish objects, we usually search for simple rules that can be applied to many observations (Occam's razor; compression). The use of simple rules is not a strict requirement for distinguishing objects, but a simplicity criterion is often used to determine which objects should be considered at all in science and mathematics. For example, it is possible to consider a glass standing on a table together as one object, but rather complex rules are required for describing how such an object behaves when pushing the table-part of the object. A much more simple set of rules of motion would be possible when the glass and the table are considered separate objects.

Because an object can only be identified by specifying how it differs from something else, any object can always be considered as composed of other objects. For example, it is possible to consider half of an electron as an object, as long as there is a way to distinguish one half from the other (even though the half-electron is not commonly addressed in physics, because it does not allow for a compact description of observations). A complete formal specification of an object, however, demands a complete description of all elements that compose that object. This leads to the idea of an elementary object (or set of objects) that cannot be further reduced, and from which everything else is made. However, the notion of an elementary object is problematic, because it cannot itself adhere to the definition of an object identified before. Let's consider the example of a world that consists of bits manipulated by a TM. To distinguish those bits from each other, there must be something inherent to those bits that allows an observer in this world (or the TM that computes that world) to treat them as distinct. However, if the bits have properties, then they are not elementary objects, because other even more fundamental concepts than just bits are required to specify what the bits really are. If the bits have no properties, then they cannot be distinguished or observed, and no computation can be performed with them at all. Hence, the assumption that bits are elementary objects that can be completely formally specified is self-contradictory. While here we used the example of bits as elementary concepts, the same goes for any object that is considered elementary, such as the smallest particle or set of particles in physics. Empirical findings, such as the notion of the Planck length or the absence of physical methods to identify smaller particles than those in the standard model of particle physics can, of course, not obtain the status of rigorous proof.

The stochastically-inclined mathematician might object that theories involving probabilities do not require actual distinctions of objects, but merely distributions over them. After all, distributions can be made even over uncountable infinite collections of things, such as real numbers. However, this is only possible because there exists a method (counting), to distinguish these numbers from each other. Without a method to distinguish different components of the distribution, a distribution cannot reflect any collective properties of its elements, would have no properties at all, and hence, would not be anything at all. Note that this discussion of probabilities does not directly relate to the distinction made in [3] between objectivists and subjectivists interpretations of probabilities, as they both assign probabilities to objects. Whether these objects are assigned a special status of "physical" is irrelevant here (although it might be interesting to investigate what this notion of "physical" exactly entails).

The assumption that there are irreducible elementary objects fits with the empiricist point of view that treats the objectivity of experience (that *objects* are observed) as given. However, our critical reflection has revealed that it is not the world-in-itself that is made of distinguishable objects, but that a subjective observer must *determine* the objects it observes or thinks about through regulative principles. Hence, it is not some (computational) structure of the world that determines our experience; instead we shape our experience through regulative principles, whose form can be be expressed in logic and mathematics. The attempt to fully formalize our experience and knowledge through the assumption that the world-in-itself (the source of our experience) is eventually made of elementary objects contradicts the necessary assumption that an object must be distinguishable to be anything at all. This also reveals why it is tempting to assume that *bits* are elementary objects [14], since the simplest distinction that can be made is between two objects; the object and what it is not.

4 Conclusion

Claims about TAIs that consider actual beings, such as humans, robots or animals, involve the assumption that observations made by these beings can be fully formalized. This assumption entails that the world consists of a set of elementary objects (e.g., bits) that are manipulated by a UTM, and can be completely formally specified. However, our critical reflection revealed that the distinction of objects through regulative rules is a *subjective* principle we necessarily use to make sense of our observations. Since we can consider this distinction only relative to a thinking or observing subject, the distinction of objects does not apply to the world-in-itself, independent of that subject. Furthermore, the *assumption* that it is possible to fully formally specify the world as a collection of irreducible elementary objects is self-contradictory. Any serious theory of algorithmic intelligence should at least require that its assumptions are free of contradiction. We also identified the reason that we are tempted to consider the world as the result of computation and the smallest particles as two distinct bits; because our observations of objects are possible through methods that can formally only be described as computation, and because the most basic distinction we can make between objects is between two (the object and what it is not). Future TAIs could benefit from both the well-founded computational theories in [3, 10, 11], *and* a critical reflection on the objects on which computation is performed.

5 FAQ

Based on the constructive discussions we had with our colleagues and reviewers follows a list of frequently asked questions and answers.

Q: What is the new insight provided in this paper? Do the considerations in any way help me to understand or build AGIs?

A: The assumption that the world can be fully formalized is invalid. Hence, the universe is not computable, no proof about AGI interacting with the real world can be given, and no absolute test for intelligence can be performed.

Q: Why don't you just give a formal account in a couple of equations of what you claim, instead of providing a lengthy argument in natural language?

A: Of course one can question the meaning of fuzzy natural language, but even in the act of questioning there are certain assumptions one necessarily has to make. It is precisely those arguments that we expose. In our discussion about knowledge, we made clear that one can only obtain knowledge by following a set of regulative rules, such as logic, and that without such rules, nothing can be known at all. Hence, our definition of knowledge does not accept anything that cannot be formally specified. Only in such a restricted setting can we derive certainty about the limits of what can be assumed without contradiction.

Q: You argue that any claim is based on assumptions, but we all know that.

A: A striking characterization of the skeptical attitude; the dominant, and at the same time, most unsatisfactory approach in philosophy nowadays. If all claims are questionable, because they are based on assumptions, then the claim that 'all claims are based on assumptions' is also based on assumptions, and hence, might not be valid. Our approach is rooted in the branch of philosophy called 'critical' (as opposed to skeptical), which identifies and investigates the assumptions underlying those reflexive claims. We do not *argue that* any claim is based on assumptions; instead we investigate *which* assumptions are made *and necessarily have to be made* to argue or question anything at all.

Q: Is the world made of bits, and, if we can't be sure, why not just *assume* it?

A: As we have argued, the concept of an object and hence, elementary objects of which everything else is made, originates from the subject that determines the objects it considers through regulative principles. This explains the logical and mathematical structure we find in our observations. We also argued that the idea that the world consists in itself of elementary objects is self-contradictory. Any serious scientific theory should at least require that its assumptions are free of contradiction.

Q: Isn't there plenty of evidence from psychology and neuroscience that human knowledge is fundamentally different from knowledge computable by a Turing machine?

A: The concept of a universal Turing machine defines the formal rules for every statement that can result from a successive chain of arguments, including the observations, objects, concepts and mechanisms described in the empirical sciences. It is therefore also a model for all the empirical sciences.

Q: Why do you introduce a *definition* of the concept of an object. Isn't it obvious that objects just exist?

A: Quite obvious indeed. Until one tries to figure out how to *construct* artificial intelligence that reasons about objects.

Q: Why don't you just ground everything in raw data (bits), rather than considering only relations without grounding? The assumption that, at the lowest level, the world consists of bits feels like an acceptable philosophical commitment.

A: The approach that starts with the objectivity of raw data is known as empiricism. However, a philosophy that is grounded in any*thing* at all, must first investigate the methods for determining the objects on which it is grounded, and hence must start with the subject that determines the objects, rather than the objects determined by the subject. The rules used to determine objects also specify the limit of what can possible be observed and known about objects. Any claims beyond those rules about the world-in-itself, can never apply to actual objects and will easily lead to self-contradictory statements, such as the existence of a elementary objects.

Q: Isn't the idea that *we* determine the objects of experience through regulative principles a strong form of solipsism (the idea that only the self really exists)?

A: No. The solipsist claims that the origin of his experience is the self. Our argument that it is strictly impossible to know the origin of our experience is precisely *the* argument against solipsism.

Q: What is the relation between your definition of knowledge as 'anything that can be positively argued about objects following the rules of a formal system', and the quite common view that knowledge is 'justified true belief'?

A: The latter definition does not specify *how* to do the justification, what the criterion for truth is, and most importantly, what the knowledge is *about*. Our approach defines the concept of an object (as anything that can be distinguished from what it is not), the method for justification (as computation), and hence, is a formally rigorous specification of *how* to justify beliefs. We also think that 'truth' should be reserved for purely abstract reasoning; there can be no truth or proof about empirical objects.

Q: What if we had a formal specification of the concept of 'human being' that involved all the possible configurations of molecules, atoms, electrons, etc, all the way down to the smallest elements that make up our world, for example, bits, would it not be possible to proof that a robot will never harm a human being?

A: Such an argument might indeed be possible if it were possible to fully formalize objects (such as a human being, or the particles in the standard model), or a distribution over them. However, as we have shown, the idea of the smallest particle cannot correspond to an empirical object, so no actual object could ever be completely formally specified.

References

[1] K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. *Monatshefte für Mathematik und Physik*, 38:173–98, 1931.

[2] J. Hernández-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539, 2010.

[3] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2004.

[4] Immanuel Kant. *Kritiek der reinen Vernunft.* Johann Friedrich Hartknoch, Riga, Zweite Originalausgabe edition, 1787.

[5] B. Jowett. *The Dialogues of Plato in Five Volumes.* Oxford University Press, Oxford, 1892.

[6] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

[7] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9:265–266, 1973.

[8] Rene Descartes. *Principia Philosophiae.* Louis Elzevir, Amsterdam, 1644.

[9] J. Schmidhuber. A computer scientist's view of life, the universe, and everything. *LNCS*, 1337:201–288, 1997.

[10] J. Schmidhuber. Ultimate cognition *à la* Gödel. *Cognitive Computation*, 1(2):177–193, 2009.

[11] B. R. Steunebrink and J. Schmidhuber. A family of Gödel machine implementations. *In this volume*, 2011.

[12] M. Tegmark. The mathematical universe. *Foundations of Physics*, 38:101–150, 2008.

[13] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–265, 1937.

[14] J. A. Wheeler. Information, physics, quantum: The search for links. In *Complexity, Entropy, and the Physics of Information*, pages 3–28. Addison-Wesley, 1990.

[15] K. Zuse. *Rechnender Raum.* Friedrich Vieweg & Sohn, Braunschweig, 1969.