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, 13, 14] 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., 8, 9, 15, 18], these theories have led to claims about intelligence in any agent [e.g., 5]. 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

J. Schmidhuber, K.R. Thórisson, and M. Looks (Eds.): AGI 2011, LNAI 6830, pp. 394–400, 2011.
© Springer-Verlag Berlin Heidelberg 2011

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.

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, 13], 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, [16]), 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, 8], which amounts to "shortest to describe", or "fastest to compute". 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 [17]), 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 Knowledge

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, 7].

We start our investigations from the question how we could convince someone or even ourselves that we know something. If we claim something, we always have to assume that to claim anything at all, means to *limit* the things we say, and that our successive claims must maintain and further specify that limitation. Although we provide a more detailed argument for making this particular assumption in [6], here it suffices to say that it is a necessary assumption, because claiming the opposite already presumes the very same assumption. To allow for a successive chain of arguments that limit what can be said, we need rules that regulate what can be said without leading to contradiction (which would cancel our previous limitations). Such systems of rules 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.

3.2 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 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. 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 determining 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 dis-

tinguish 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 compact descriptions 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 (and 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.

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 [17], 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, 10–14], *and* a critical reflection on the objects on which computation is performed.

References

[1] Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. Monatshefte für Mathematik und Physik 38, 173–198 (1931)
[2] Hernández-Orallo, J., Dowe, D.L.: Measuring universal intelligence: Towards an anytime intelligence test. Artificial Intelligence 174(18), 1508–1539 (2010)
[3] Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2004)
[4] Kant, I.: Kritiek der reinen Vernunft. Johann Friedrich Hartknoch, Riga, Zweite Originalausgabe edition (1787)
[5] Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Minds and Machines 17(4), 391–444 (2007)
[6] Pape, L., Kok, A.: Real-world limits to algorithmic intelligence (2011), Online version: `http://www.idsia.ch/~pape/papers/pape2011agilong.pdf`
[7] Descartes, R.: Principia Philosophiae. Louis Elzevir, Amsterdam (1644)
[8] Schmidhuber, J.: A computer scientist's view of life, the universe, and everything. In: Freksa, C., Jantzen, M., Valk, R. (eds.) Foundations of Computer Science. LNCS, vol. 1337, pp. 201–288. Springer, Heidelberg (1997)
[9] Schmidhuber, J.: Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. International Journal of Foundations of Computer Science 13(4), 587–612 (2002)
[10] Schmidhuber, J.: The Speed Prior: a new simplicity measure yielding near-optimal computable predictions. In: Kivinen, J., Sloan, R.H. (eds.) COLT 2002. LNCS (LNAI), vol. 2375, pp. 216–228. Springer, Heidelberg (2002)

[11] Schmidhuber, J.: Completely self-referential optimal reinforcement learners. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 223–233. Springer, Heidelberg (2005)
[12] Schmidhuber, J.: Gödel machines: fully self-referential optimal universal self-improvers. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 199–226. Springer, Heidelberg (2006); Variant available as arXiv:cs.LO/0309048
[13] Schmidhuber, J.: Ultimate cognition à la Gödel. Cognitive Computation 1(2), 177–193 (2009)
[14] Steunebrink, B.R., Schmidhuber, J.: A family of Gödel machine implementations. In: Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.) AGI 2011. LNCS(LNAI), pp. 268–273. Springer, Heidelberg (2011)
[15] Tegmark, M.: The mathematical universe. Foundations of Physics 38, 101–150 (2008)
[16] Turing, A.M.: On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society 42, 230–265 (1937)
[17] Wheeler, J.A.: Information, physics, quantum: The search for links. In: Complexity, Entropy, and the Physics of Information, pp. 3–28. Addison-Wesley, Reading (1990)
[18] Zuse, K.: Rechnender Raum. Friedrich Vieweg & Sohn, Braunschweig (1969)