sciendo

Journal of Artificial General Intelligence

Special Issue "On Defining Artificial Intelligence"
—Commentaries and Author's Response

Volume 11, Issue 2
February 2020
DOI: 10.2478/jagi-2020-0003
Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Contents

Journal of Artificial General Intelligence 11(2) 1-4, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2020-01-16
Accepted 2020-01-16

Introduction to the JAGI Special Issue "On Defining Artificial Intelligence" —Commentaries and Author's Response

Dagmar Monett DAGMAR.MONETT@AGISI.ORG
Berlin School of Economics and Law, and
AGISI.org
Berlin, Germany

Colin W. P. Lewis COLIN.LEWIS@AGISI.ORG
AGISI.org
Warsaw, Poland

Kristinn R. Thórisson THORISSON@RU.IS
Department of Computer Science, Reykjavik University,
and Icelandic Institute for Intelligent Machines
Reykjavik, Iceland

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Pei Wang's paper titled "On Defining Artificial Intelligence" was published in a special issue of the Journal of Artificial General Intelligence (JAGI) in December of last year (Wang, 2019). Wang has been at the forefront of AGI research for over two decades. His non-axiomatic approach to reasoning has stood as a singular example of what may lie beyond narrow AI, garnering interest from NASA and Cisco, among others. We consider his article one of the strongest attempts, since the beginning of the field, to address the long-standing lack of consensus for how to define the field and topic of artificial intelligence (AI). In the recent AGISI survey on defining intelligence (Monett and Lewis, 2018), Pei Wang's definition,

> *The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets "intelligence" as a form of "relative rationality"* (Wang, 2008),

was the most agreed-upon definition of artificial intelligence with more than 58.6% of positive ("strongly agree" or "agree") agreement by the respondents (N=567).

Due to the greatly increased public interest in the subject, and a sustained lack of consensus on definitions for AI, the editors of the Journal of Artificial General Intelligence decided to organize a special issue dedicated to its definition, using the target-commentaries-response format. The goal of this special issue of the JAGI is to present the commentaries to (Wang, 2019) that were received together with the response by Pei Wang to them.

A total of 110 leading experts (31.8% female, 68.2% male) were invited to contribute with commentaries to the target article. The criteria for selection considered a conjunction of research in AI and AGI related topics, scientific work on defining AI as a field or as a concept, (co-)authorship of international and national AI-related reports, (co-)authorship of books on AI, as well as chair activities in major AI conferences, among other criteria.

More than 1300 email messages including invitations, several follow-ups and reminders per invited expert, as well as organisational emails exchanged in all phases of the editorial process, were sent. The deadline for submission was extended several times upon some authors requests.

42 experts (38.2%) rejected the invitations explicitly. 48 experts (43.6%) didn't respond to our call.[1] Other general statistics are presented in Table 1.

		Female		**Male**		
Invites ...	**No.**	*% of total*	*% of female*	*% of total*	*% of male*	**Total**
... sent	110	35		75		110
		31.8	100.0	68.2	100.0	100.0%
... accepted	20	0		20		20
		0.0	0.0	18.2	26.7	18.2%
... rejected	42	16		26		42
		14.5	45.7	23.6	34.7	38.2%
... with no answer back	48	19		29		48
		17.3	54.3	26.4	38.7	43.6%

Table 1: Some general statistics of the editorial process regarding invitations to contribute.

We received twenty commentaries, those by Joscha Bach, Gianluca Baldassarre and Giovanni Granato, Istvan Berkeley, Francois Chollet, Matthew Crosby and Henry Shevlin, John Fox, John Laird, Shane Legg, Peter Lindes, Tomas Mikolov, William J. Rapaport, Raúl Rojas, Marek Rosa, Roger C. Schank, Aaron Sloman, Peter Stone, Richard S. Sutton, Kristinn R. Thórisson, Alan Winfield, and Roman V. Yampolskiy. All commentaries were accepted after peer-review.

If the reader was expecting a consensus around defining AI, we are afraid we have to disappoint them. We have received many kinds of responses: commentators that don't agree with Pei Wang's definition and provide their own, those that don't consider we need new definitions at all, those that agree with Wang's but still provide a new definition of AI, as well as those that additionally prefer to comment about other topics they feel are also important. A very colored spectrum around defining the most important concept of the AI field!

The commentaries published in this special issue are grouped in four parts:

- **Part I** includes one introductory commentary by Kristinn R. Thórisson (2020) that addresses central aspects of the target article from the editors' point of view.

- **Part II** contains sixteenth invited peer commentaries (Bach, 2020; Baldassarre and Granato, 2020; Berkeley, 2020; Chollet, 2020; Crosby and Shevlin, 2020; Fox, 2020; Laird, 2020;

1. Most striking in these numbers is the glaring absence of female authors. A common reason among female academics for rejecting our invitation to contribute was *overcommitment*. As a community, we may want to think of new, different ways of engaging the full spectrum of AI practitioners if we value inclusion as an essential constituent of a healthy scientific growth. Self determination and willingness to participate are also essential.

Legg, 2020; Lindes, 2020; Mikolov, 2020; Rapaport, 2020; Rojas, 2020; Rosa, 2020; Stone, 2020; Sutton, 2020; Yampolskiy, 2020) that address the target article explicitly, alphabetically ordered with respect to the surname of their first contributors.

- **Part III** continues with Pei Wang's response (Wang, 2020) to those invited commentaries that are included in Part II.
- **Part IV** finishes this especial issue of the JAGI. It presents other three invited peer commentaries (Schank, 2020; Sloman, 2020; Winfield, 2020) that address other general topics related to the target article, like defining artificial intelligence, but that do not necessarily refer to it explicitly.

We are convinced that a variety of opinions on defining AI, especially as seen through the spectacles of a group of leading AI authorities, will be remarkably influential both for the field and for defining machine intelligence.

We trust that this special issue of the JAGI will become a transcending referent on defining AI and that, in Pei Wang's words (Wang, 2020), it will constitute the beginning, not the ending, of that discussion.

Acknowledgments

We want to thank all authors for their time, their commitment, and the high quality of their contributions.

References

Bach, J. 2020. When Artificial Intelligence Becomes General Enough to Understand Itself. Commentary on Pei Wang's Paper "On Defining Artificial Intelligence". *Journal of Artificial General Intelligence* 11(2):15–18.

Baldassarre, G. and Granato, G. 2020. Goal-Directed Manipulation of Internal Representations Is the Core of General-Domain Intelligence. *Journal of Artificial General Intelligence* 11(2):19–23.

Berkeley, I. 2020. AI: A Crowd-Sourced Criterion. A Commentary on Pei Wang's Paper "On Defining Artificial Intelligence". *Journal of Artificial General Intelligence* 11(2):24–26.

Chollet, F. 2020. A Definition of Intelligence for the Real World? *Journal of Artificial General Intelligence* 11(2):27–30.

Crosby, M. and Shevlin, H. 2020. Defining Artificial Intelligence: Resilient Experts, Fragile Geniuses, and the Potential of Deep Reinforcement Learning. *Journal of Artificial General Intelligence* 11(2):31–34.

Fox, J. 2020. Towards a Canonical Theory of General Intelligence. *Journal of Artificial General Intelligence* 11(2):35–40.

Laird, J. 2020. Intelligence, Knowledge & Human-like Intelligence. *Journal of Artificial General Intelligence* 11(2):41–44.

Legg, S. 2020. A Review of "On Defining Artificial Intelligence". *Journal of Artificial General Intelligence* 11(2):45–46.

Lindes, P. 2020. Intelligence and Agency. *Journal of Artificial General Intelligence* 11(2):47–49.

Mikolov, T. 2020. Why Is Defining Artificial Intelligence Important? *Journal of Artificial General Intelligence* 11(2):50–51.

Monett, D. and Lewis, C. W. P. 2018. Getting clarity by defining Artificial Intelligence–A Survey. In Müller, V. C., ed., *Philosophy and Theory of Artificial Intelligence 2017*, volume SAPERE 44. Berlin: Springer. 212–214.

Rapaport, W. J. 2020. What Is Artificial Intelligence? *Journal of Artificial General Intelligence* 11(2):52–56.

Rojas, R. 2020. On Pei Wang's Definition of Artificial Intelligence. *Journal of Artificial General Intelligence* 11(2):57–59.

Rosa, M. 2020. On Defining Artificial Intelligence—Commentary. *Journal of Artificial General Intelligence* 11(2):60–62.

Schank, R. C. 2020. What Is AI? *Journal of Artificial General Intelligence* 11(2):89–90.

Sloman, A. 2020. A Philosopher-Scientist's View of AI. *Journal of Artificial General Intelligence* 11(2):91–96.

Stone, P. 2020. A Broader, More Inclusive Definition of AI. *Journal of Artificial General Intelligence* 11(2):63–65.

Sutton, R. S. 2020. John McCarthy's Definition of Intelligence. *Journal of Artificial General Intelligence* 11(2):66–67.

Thórisson, K. R. 2020. Discretionarily Constrained Adaptation Under Insufficient Knowledge and Resources. *Journal of Artificial General Intelligence* 11(2):7–12.

Wang, P. 2008. What Do You Mean by "AI"? In Wang, P., Goertzel, B., and Franklin, S., eds., *Artificial General Intelligence 2008. Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications*, volume 171. Amsterdam, The Netherlands: IOS Press. 362–373.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Wang, P. 2020. On Defining Artificial Intelligence—Author's Response to Commentaries. *Journal of Artificial General Intelligence* 11(2):73–86.

Winfield, A. 2020. Intelligence Is Not One Thing. *Journal of Artificial General Intelligence* 11(2):97–100.

Yampolskiy, R. V. 2020. On Defining Differences Between Intelligence and Artificial Intelligence. *Journal of Artificial General Intelligence* 11(2):68–70.

Part I

Introductory commentary

Journal of Artificial General Intelligence 11(2) 7-12, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-22
Accepted 2019-11-11

Discretionarily Constrained Adaptation Under Insufficient Knowledge & Resources

Kristinn R. Thórisson THORISSON@RU.IS
Department of Computer Science, Reykjavik University,
and Icelandic Institute for Intelligent Machines
Reykjavik, Iceland

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

In his paper "On Defining Artificial Intelligence" Pei Wang (2019) defines intelligence as "adaptation with insufficient knowledge and resources." This highly compact definition of a term used to name a field of research, as well as some of its products, cuts to the heart of the natural phenomenon we call "intelligence" by addressing an issue that I will paraphrase as *autonomous handling of novelty*. More on this below.

Wang points out—and rightly so—that definitions affect the way phenomena get studied in science. He also points out the side effect of *premature definitions*: They can lead us astray. Before we have a good scientific understanding of a particular phenomenon it is however next to impossible to come up with a good scientific definition—how could we possibly define something properly that we don't understand well? And yet, to study any phenomenon scientifically requires making *some* assumptions about that phenomenon, especially its relation to better-understood ones. How can this conundrum be addressed?

1. Definitions Affect The Way Phenomena Are Studied

In the early days of any research field we rely on "working definitions"—so called to remind us that they should be improved as soon as possible (and not sooner!). Any good definition captures the essence of a phenomenon it targets when that phenomenon is well understood; a good *working* definition cannot do so, since the subject is not understood. Then what use is it? Actually it is rather important, but not for the same purpose as for definitions that are produced in the later phases of a research field, after the subject matter is better understood. Rather, the reason working definitions are important is because of their ability to help researchers focus on critical issues and *key aspects* of the phenomenon under scrutiny: While a penultimate definition's job is to give the full and complete picture of the thing it refers to in the shortest amount of space, a working definition serves a related but slightly different role as a searchlight: It should put key aspects of the phenomenon center of stage. Long working definitions are thus often preferable to shorter ones, especially for complex, intricate and integrative phenomena like the ecosystem, society, and mind. The urge to simplify, often through too-compact a definition, risks lopping off aspects that are not only important but *integral* to the very phenomenon of interest (Thórisson, 2013). To take some illustrative examples we might mention for instance *contaminants* in forensic investigations, *time* in developmental studies, *weather* and *biochemistry* in ecological studies—which, if left out, would

significantly affect the way research was conducted, impeding progress for decades, even centuries. If a key aspect of a phenomenon under scientific study is forgotten in a working definition, we may in effect unknowingly be redefining our subject and, from that point onward, be studying a completely different phenomenon! Our findings, theories and data may in this case only partially generalize, or perhaps not at all, to the original phenomenon of interest. This danger is greater for highly intricate, non-linear systems than for simpler ones. To take an example, researchers in the field of developmental psychology aim to unravel the nature of how the cognitive control mechanisms of individuals change over years and decades. If they were to use a working definition of cognitive development along the lines of "the difference in abilities of the same individual between two points in time" they would be emphasizing correlation over progression: Instead of helping researchers approach cognitive growth as an architectural process influenced by the mind's interaction with the environment, this definition would draw them towards point measurements and statistical comparisons; towards oversimplification. Looking for principles of morphing cognitive architectures this way would be futile, or at best extremely slow: Leaving out a defining part of a new research field's central phenomenon does not bode well for scientific progress.

2. Novelty Demands Generality

When defining artificial intelligence (AI), the term "artificial" has never been under scrutiny: It simply means "made by people." The second part, "intelligence," being a very useful term in the vernacular, is a polysemous term for a phenomenon that originated in nature and begs to be named: The ability of animals to solve problems, learn and create new things, communicate, reason, and many other things. In fact, there seem to be so many things relevant to the phenomenon of intelligence that by the end of the last decade AI researchers had come up with over 28 (working) definitions (cf. (Legg and Hutter, 2007; Monett and Lewis, 2018)), a number that undoubtedly has grown since.

Defining AI is thus in large part synonymous with the task of defining intelligence. Natural intelligence is specialized to handle problems in the physical world; artificial intelligence targets problems chosen by its creators. Instances of either can be placed somewhere along a dimension of *generality*, as defined by an agent's ability to handle variety, complexity, and novelty. Incidentally, when we say "handle" we mean the ability of an agent to achieve goals with respect to its targeted purpose and deal with the many things it encounters, as well as explain, predict, and even re-create them (as models, or in some other form) *autonomously*, that is, without "calling home" (cf. (Thórisson et al., 2016; Thórisson and Helgason, 2012)). The interactions between the myriad of relevant variables encountered by any agent operating in the physical world, even just walking through a city for one hour, is enormously large—so gigantic that there is no way to precompute it all and store in a lookup table, should we be foolish enough to try: For all practical purposes the physical world presents *novelty at every turn* that must be dealt with on-demand.

The concept of 'novelty' is of course a gradient: The scenarios we encounter every day may be in many ways similar to the ones we encountered yesterday, but they are never identical down to every detail. And sometimes they are *very different*. But due to the impossibility of defining everything up front, and knowing everything beforehand, both natural and artificial intelligences must rely on creativity and learning as a major way to operate in the physical world. Because the world presents this wealth of novelty, we are constantly in a state of lacking knowledge. The purpose of intelligence is to figure out what knowledge is needed, produce that knowledge by any

means necessary, and allow us to move on. This is the way both natural and artificial intelligences can handle novel problems, tasks, goals, environments and worlds. No task takes zero time or energy to perform—and neither does thinking. Time and energy present additional constraints on this effort, and cannot be removed. Addressing these challenges is what all intelligent agents must be able to do, as well as any and all other constraints that may come their way. This is why we continue to push our machine's abilities increasingly towards the 'general' end of that spectrum.

3. Intelligence Means Figuring Things Out

A key feature of human intelligence, in contrast to special algorithms, is its ability to generate novel sequences of actions, events, thoughts, etc.—programs—that bridge from idealized models of the world to physical actions that affect and change the world. For three quarters of a century we have known how to make electronic computers effectively run predefined programs, but we still don't know how to make machines that can *create novel programs* effectively.

This capability is nevertheless what environmental novelty necessitates, and thus quite possibly the single defining feature that no other phenomenon than intelligence can make claim to. So it could—and possibly should—be an integral part of a definition of intelligence. This is what Pei Wang's definition does so elegantly and why 'limited knowledge and resources' is at the center of his definition. Is he saying that humans, because they are intelligent, never do anything by rote memory or by routine? Not at all. Is he saying that under no circumstances do people have *sufficient* knowledge and resources? I don't think so. He is pointing out that if that was *all* they did, they'd hardly be called intelligent; and that the other aspect of what they routinely do, and *must* do—*figure out stuff*—is what makes them unique and unlike any other process—worthy of the label 'intelligence.' Wang has cleverly isolated a key aspect of (general) intelligence that many others have overlooked or completely excluded: The ability—and unavoidability—of intelligent agents operating under insufficient knowledge and resources to *necessarily generate new programs*.

So, with the 'assumption of insufficient knowledge and resources' (a.k.a. AIKR) Wang boils down the definition of AI to this particular constant activity of intelligence: To innovate, to try to figure things out, in light of resource scarcity. What are the 'resources' that are being referred to here? Information, planning time, sensing time, reasoning time, etc.—anything that may be partial or missing when people are faced with new things. By focusing on this small but hugely important aspect of the numerous things that (general) intelligences can do, and that he could have focused on but chose not to, Wang brilliantly highlights the one thing that *must* call for some sort of *generality*—the ability of a single agent to handle the *unknown variety of the world* throughout its lifetime.

4. Adaptation Through Reasoning

The first term in Wang's definition is "adaptation," a term that is quite a bit less specific than the rest of his definition. The concept of adaptation is well known in the context of evolution, where it refers to processes that change in light of external forces (c.f. (Holland, 1975)). It is also used for much simpler processes such as sand that "adapts" to a bucket's form factor when it's poured in. This is hardly what Wang means. But what about the evolutionary sensebiological adaptation? Here the term refers to both the genotype and the phenotype, as they change in response to the evolutionary process of survival of the fittest. I would argue that the sense of "adaptation" called for by Wang's

definition is also quite different from this, in some fundamental ways. So could his definition be improved, still?

Thought does not seem to "adapt" to "forces" in any way similar to genetic mechanisms: Evolution "blindly" follows a relatively simple algorithm that generates lots of variation (individuals) and is left with "whatever sticks;" thought, in contrast, relies on *reasoning*: A systematic application of logic to models built from experience. A result of this, and a clear indication at that, is the fact that any generally intelligent agent worth its salt can *explain* important aspects of its knowledge—*what* it does, *why*, and *how*. Evolutionary processes can absolutely not do this, because they cannot be given arbitrary goals. The term "adaptation" requires thus, in my opinion, additional clarification and qualification.

Reasoning plays an important role in intelligence not because it is exclusively human (it isn't; cf. (Balakhonov and Rose, 2017)) but because it is necessary for cumulative learning (Thórisson et al., 2019): Due to the AIKR there will simply be far too many things and options worthy of inspection and consideration, for any intelligent agent operating in the physical world. When building up coherent and compact knowledge through experience, through cumulative learning, reasoning processes ensure that prior experience can be used to make sense of the new, by e.g. eliminating improbable or ridiculous hypotheses about them (e.g. we can dismiss the claim of a rollerblade vendor that their product "enables its user to go through walls," before we see their rollerblades—and even if we didn't know what rollerblades are, because we consider the rules "solid objects cannot go through each other" and "footwear is unlikely to affect the solidity of its user" to be stronger, especially in light of our well-supported experience that *nothing* can affect the solidity of *anything* in that way). There is no denying that intelligence *requires* the ability to create sensible goals and use reasoning to manage them—goals define what is accepted and not accepted when addressing some task, environment, or problem; by specifying their *constraints*. Goals are thus a kind of temporally-bounded requirement on intelligence, and trying to create a generally intelligent machine that does not have this ability seems tautological.

5. Knowledge-Scarce Sense-Making

If nature is "the blind watchmaker," thought is the "partially-informed sense-maker": Based on an agent's changing needs and wishes relative to its environment, an agent forms multiple (explicit or implicit) sub-goals, which it uses in combination with reasoning to cumulatively build up a collection of reliable and actionable knowledge, to predict, achieve goals, explain, and re-create the phenomena that it models from experience (Bieger, Thórisson, and Steunebrink, 2017; Thórisson et al., 2016). A closely related hallmark of (general) intelligence is thus an ability to freely define, compare, and change goals: Other things being equal, increased flexibility in this direction means a greater ability to solve problems, classify concepts, create things, analyze the world and one's own thoughts.

Since both biological processes and intelligent agents can be said to "adapt" to their environment, albeit in different ways, the term chosen to address this aspect of intelligence should help separate these two different meanings clearly. We can either use a different term to 'adaptation', or qualify it further. I propose to extend Pei Wang's otherwise excellent definition, to include the following: *Intelligence is discretionarily constrained adaptation under insufficient knowledge and resources*.

What does this mean? Simply that the adaptation may be arbitrarily constrained at the discretion of the agent itself or someone/something else. This clearly separates this use of 'adaptation' from its sense in the context of natural evolution, whose course is determined by uniform physical laws. To be called intelligent, in contrast to evolution, the adaptation in question needs to have a capacity to handle arbitrary constraints of many forms, including "doing the dishes without breaking them" as well as "doing the dishes before noon." It also must be capable of inventing such constraints in light of multiple (often conflicting) goals, e.g. "grading student assignments before noon frees up the afternoon for paper writing." Constraining the adaptation 'discretionarily' means that constraints can be freely added to the way the adaptation is allowed to proceed, in ways that are independent of the nature of the task, environment, goal, or problem—that the specification of the "space of acceptable adaptation" can be limited at the problem designer's discretion as well as the agent's.

6. What It All Means

For all the reasons presented above I consider Pei Wang's definition of intelligence the most important one proposed to date. Unlike virtually all other existing definitions it "carves out" the very thing that is unique about intelligence. Let's not forget, however, that it's a *working* definition, which means it should be improved—soon. My addition is not intended to change it, only to constrain it in a way that I consider important for its purpose as a working definition: To help us focus on a core aspect of intelligence while reducing the chance of misinterpretation by separating it more clearly from alternative interpretations.

What may be the relevance of this working definition for the field of AI? Well, it proposes to put an issue front and center that has never really been at the center of our approach to intelligence before (except in Pei Wang's own writings; cf. (Monett and Lewis, 2018; Wang, 2006)). This has far-reaching implications which can be viewed from several angles; let us conclude by taking a brief look at one. This definition clears up the apparent rift between ready-made software systems and those that are truly intelligent: According to Wang, traditional software programs are not intelligent because they cannot create new programs. Clarifying this is actually good for the field, even though many may raise an eyebrow or two, and possibly even make some really mad, because historically the field has spent too much time and effort in discussing whether this or that program is ("truly") intelligent—programs that, besides their application and data, when it comes down to it, were difficult to distinguish in any way, shape, form or function from standard software. The definition puts creativity right alongside intelligence itself, which also makes a lot of sense: What would a super-smart intelligence without creativity look like? Seems like an oxymoron to me. A clear sign of the immaturity in any research field is the number of unexplained contradictions. One of these is the so-called "AI effect," whereby some "AI solutions"—diligently pursued under the AI banner for years or decades—become "just algorithms" when they (inevitably) are adopted by mainstream computer science. Wang's definition explains the source of this "effect": Software systems that can be produced through the traditional allonomic principles of software development (cf. (Thórisson, 2012)), and run according to the same principles, are simply *software*—no amount of wishful thinking will make them "intelligent." They may mirror some (small) aspect of human and animal intellect, but they lack a central feature: *Discretionarily constrained adaptation under insufficient knowledge and resources*. For building a truly intelligent software system, traditional software development methods will not suffice; additional principles are required that have to do with intelligence proper, namely, the central theme of this fundamentally new definition.

References

Balakhonov, D. and Rose, J. 2017. Crows Rival Monkeys in Cognitive Capacity. *Nature Sci. Rep.* 7(8809):1–8.

Bieger, J., Thórisson, K. R., and Steunebrink, B. 2017. Evaluating understanding. In *Proceedings of the IJCAI Workshop on Evaluating General-Purpose AI*.

Holland, J. 1975. *Adaptation in natural and artificial systems*. University of Michigan Press.

Legg, S. and Hutter, M. 2007. A Collection of Definitions of Intelligence. In Goertzel, B. and Wang, P., eds., *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, volume 157, 17–24. UK: IOS Press.

Monett, D. and Lewis, C. W. P. 2018. Getting clarity by defining Artificial Intelligence–A Survey. In Müller, V. C., ed., *Philosophy and Theory of Artificial Intelligence 2017*, volume SAPERE 44. Berlin: Springer. 212–214.

Thórisson, K. R. and Helgason, H. P. 2012. Cognitive architectures & autonomy: A comparative review. *Journal of Artificial General Intelligence* 3(2):1–30.

Thórisson, K. R., Kremelberg, D., Steunebrink, B. R., and Nivel, E. 2016. About Understanding. In Steunebrink, B. R., Wang, P., and Goertzel, B., eds., *Artificial General Intelligence (AGI-16)*, 106–117. New York, USA: Springer-Verlag.

Thórisson, K. R., Beiger, J., Li, X., and Wang, P. 2019. Cumulative Learning. In Hammer, P., Agrawal, P., Goertzel, B., and Iklé, M., eds., *Artificial General Intelligence (AGI-19)*, 198–209. Shenzhen, China: Springer-Verlag.

Thórisson, K. R. 2012. A new constructivist AI: From manual construction to self-constructive systems. In Wang, P. and Goertzel, B., eds., *Theoretical Foundations of Artificial General Intelligence*. Atlantis Thinking Machines. 145–171.

Thórisson, K. R. 2013. Reductio ad Absurdum: On Oversimplification in Computer Science and its Pernicious Effect on Artificial Intelligence Research. In Abdel-Fattah, A. H. M. and Kuhnberger, K.-U., eds., *Proceedings of the Workshop Formalizing Mechanisms for Artificial General Intelligence and Cognition (Formal MAGiC)*, 31–35. Osnabrück: Institute of Cognitive Science.

Wang, P. 2006. Artificial Intelligence: What It Is, And What it Should Be. In Lebiere, C. and Wray, R., eds., *Papers from the AAAI Spring Symposium on Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*. 97–102.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Part II

Invited peer commentaries

Journal of Artificial General Intelligence 11(2) 15-18, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-09
Accepted 2019-11-05

When Artificial Intelligence Becomes General Enough to Understand Itself. Commentary on Pei Wang's paper "On Defining Artificial Intelligence"

Joscha Bach JOSCHA@AIFOUNDATION.COM
AI Foundation
San Francisco, California, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

When the field of Artificial Intelligence was founded, John McCarthy (Minsky et al., August 31 1955) described it as *"to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. ... For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving."* However, with respect to a definition of intelligence, AI researchers arguably found themselves to be in a similar position as early biologists with respect to a definition for life. Like life, intelligence had a referent by example (that is, human intelligence, and to a lesser degree cognitive abilities of complex animals), but not a clear functional definition yet. Attempts to define life and separate it from non-life hinged on enumerating properties, like growth, reproduction, metabolism and adaptation. While cells were already discovered in the 17th century (Hooke, 1665), it took until 1839 before cell theory fundamentally explained life as the functional dynamics of cells: all living organisms are composed of one or more cells, which are their basic unit of structure and organization, and arise from earlier cells (Schleiden, 1839; Schwann, 1839). Biology could not start out with this functional definition, because it had to be uncovered in the course of paradigmatic research in the discipline, which had to start out with working definitions that got revised as understanding progressed.

Similarly, Pei Wang (2019) distinguishes a *dictionary definition* and *working definitions* for AI, the former serving as a reference to the field, and the latter describing a research paradigm that is meant to scale into a functional model. I think this implies that Artificial Intelligence research has to concern itself with studying the nature of intelligence. If it succeeds and identifies its subject, a lot of the working definitions will either disappear or turn out to relate to aspects of the same subject, and be replaced by a functional one. Obtaining a functional definition of intelligence is intimately related to succeeding in building an Artificial Intelligence (in the sense of AGI): the definition will probably turn out to be identical to a general specification for implementing such a system. In Wang's sense, a working definition amounts to or at least points into the direction of a hypothesis on how to realize AI.

In the past, various attempts at defining AI in such terms were made. For instance, starting from an understanding of intelligence as universal problem solving, Newell, Shaw and Simon proposed the *general problem solver* (GPS; Newell, Shaw, and Simon (1959)). The GPS did not succeed beyond simple problems like the Tower of Hanoi task, because it ran into a combinatorial explosion when applied to real-life problems. In response, Laird and Newell suggested to implement problem

solving using a library of cognitive skills (Laird and Newell, 1983), which led to the development of the cognitive architecture Soar (Laird, Newell, and Rosenbloom, 1987). By extending the set of tasks that Soar could tackle, its authors hoped to cross the threshold into a generally intelligent system.

A different line of thinking started from Irving John Good's notion of intelligence an ability for self improvement (Good, 1965). Good's paper is often considered as the origin of the concept of an *AI singularity* (i.e. a catastrophic take over by a runaway self improving super intelligence), though arguably, this idea has been described earlier in Science Fiction literature (e.g. Leiber (1962)). The idea of intelligence as self improvement has lead to Jürgen Schmidhuber's Gödel Machine (Schmidhuber, 2003), a system that performs provably optimal self improvements. The realization of a Gödel machine will however require the implementation of a system that can perform such proofs efficiently (Steunebrink and Schmidhuber, 2012). Unless there are intrinsic limits to the intelligence of any physical system, and human intelligence is already close to this limit, the idea of self improving AI implies that such a system is going to dramatically surpass our own abilities. Pei Wang does indeed believe that human intelligence is close to the limit of that of any possible intelligent system (Wang, Liu, and Dougherty, 2018), although he accepts that the capacity for self modification is an important part of intelligence. (Confusingly and in my view needlessly, Wang defines computation as *"predetermined algorithm to realize a function that maps input data to output data"* and takes that to mean that this algorithm cannot modify itself, which would imply that computation defines *"a constant and invariant ability,"* so that a computational system cannot be intelligent.)

Another AI research program starts out with Ray Solomonoff's insight (Solomonoff, 1964) that an intelligent agent ought to find the shortest among the models that best predict present observations from past observations, for all observations. Together with a notion of reward maximization, *Solomonoff induction* leads to Marcus Hutters AIXI paradigm for universal intelligence (Hutter, 2005). While Wang points out that AIXI is itself not computable, numerous works aim at efficiently approximating AIXI agents (e.g. Legg (2008); Franz et al. (2018). Last but not least, Karl Friston's proposal of the Free Energy Principle (Friston, Kilner, and Harrison, 2006), together with the concept of self organization, proposes to understand intelligence from the perspective of minimization of free energy in an agent's environment (Friston, 2010).

Pei Wang's own working definition, *"intelligence is the capacity of an information processing system to adapt to its environment while operating with insufficient knowledge and resources,"* is dating back to 1995 and may by itself constitute the weakest part of his valuable contribution. I am especially concerned that this definition does not depend on the agent itself, but on its environment, and relies on shortcomings rather than capabilities of the agent. Why would an intelligent agent that is offered *unbounded* resources not make use of them? Why should an agent with *sufficient* knowledge be considered less intelligent? This is not empty theoretical sophistry. In my view, it is not plausible that Pei Wang's own NARS architecture should be considered suitable as a proposal for general intelligence if it is in principle unable to reach the performance of narrow AI systems when given sufficient resources. While generally intelligent systems may need additional resources to solve problems that narrow AI systems already implement, their capabilities should be a superset of those of narrow systems.

Another, slightly less concerning issue may constitute the comparison of intelligence definitions using percepts, mental states and actions ($\langle P, S, A \rangle$), which is used throughout the paper. In my view, percepts and actions cannot be readily treated as an interface to the environment. Instead, percepts

and actions are themselves representational states. An understanding of perception and action will generally not be independent of the model of intelligence of the agent itself, hence making the comparison between different approaches in this framing difficult or even impossible.

Unlike Wang, I don't think of intelligence is the ability to use *"bounded resources for unbounded problems,"* but as the ability to deal with complexity by making models, usually in the service of a complex control task (such as the persistent existence of a complex agent in an entropic universe). According to *the Good Regulator theorem* (Conant and Ashby, 1970), an effective control system needs to implement a model that is isomorphic to the system it regulates. It appears to me that a *general AI* is one that, when presented with a sufficiently complex problem, is able to come up with a model that encompasses the general conditions and preconditions of its own existence, i.e. requirements to a universe that can contain the intelligent agent, the nature of representational languages, and the implementation of the generally intelligent system itself.

In other words, the ability to create a generally AI may be a necessary and sufficient condition for general AI. In this sense, the question whether human intelligence qualifies as generally intelligent is still an open one. Our inquiry into how to build a generally intelligent system is in no small part an attempt to understand our own nature.

References

Conant, R. C. and Ashby, W. R. 1970. Every good regulator of a system must be a model of that system. *International Journal for Systems Science* 1:89–97.

Franz, A., Löffler, M., Antonenko, A., Gogulya, V., and Zaslavskyi, D. 2018. Introducing WILLIAM: a system for inductive inference based on the theory of incremental compression. In *International Conference on Computer Algebra and Information Technology*.

Friston, K., Kilner, J., and Harrison, L. 2006. A free energy principle for the brain. *J. Physiol.* 100:70–87.

Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11:127–138.

Good, I. J. 1965. Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers* 6:31ff.

Hooke, R. 1665. *Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses*. London: J. Martyn and J. Allestry.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. EATCS Book, Springer.

Laird, J. E. and Newell, A. 1983. A Universal Weak Method: Summary of results. In *IJCAI*, volume 2, 771772.

Laird, J. E., Newell, A., and Rosenbloom, P. S. 1987. SOAR: An architecture for general intelligence. *Artificial Intelligence* 33:1–64.

Legg, S. 2008. *Machine Superintelligence*. Doctoral dissertation, University of Lugano.

Leiber, F. 1962. *The Creature from Cleveland Depths*.

Minsky, M., Rochester, N., Shannon, C., and McCarthy, J. August 31, 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Newell, A., Shaw, J. C., and Simon, H. A. 1959. Report on a general problem-solving program. In *Proceedings of the International Conference on Information Processing*, 256–264.

Schleiden, M. J. 1839. Beiträge zur Phytogenesis. In *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*. 137–176.

Schmidhuber, J. 2003. Goedel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements. arXiv:cs/0309048 [cs.LO]. https://arxiv.org/abs/cs/0309048.

Schwann, T. 1839. *Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachsthum der Thiere und Pflanzen*. Berlin: Sander.

Solomonoff, R. J. 1964. A formal theory of inductive inference. *Information and Control* 7:224–254.

Steunebrink, B. R. and Schmidhuber, J. 2012. Towards an Actual Gödel Machine Implementation. In P. Wang, B. G., ed., *Theoretical Foundations of Artificial General Intelligence*. Springer.

Wang, P., Liu, K., and Dougherty, Q. 2018. Conceptions of Artificial Intelligence and Singularity. *Information* 9.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 19-23, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-11-20
Accepted 2020-01-16

Goal-Directed Manipulation of Internal Representations Is the Core of General-Domain Intelligence

Gianluca Baldassarre GIANLUCA.BALDASSARRE@ISTC.CNR.IT
Laboratory of Embodied Computational Neuroscience (LOCEN)
Institute of Cognitive Sciences and Technologies (ISTC)
National Research Council (CNR)
Roma, Italy

Giovanni Granato GIOVANNI.GRANATO@ISTC.CNR.IT
Laboratory of Embodied Computational Neuroscience (LOCEN)
Institute of Cognitive Sciences and Technologies (ISTC)
National Research Council (CNR)
Roma, Italy

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

1. A narrow definition for an AI research agenda

The target article (Wang, 2019) claims that a shared *working definition* of intelligence is useful to guide research in artificial intelligence (AI). The Author of the article thus proposes that "*intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources*." We think such definition fits more 'general-domain intelligence' than intelligence *tout court*. Our definition of intelligence is indeed wider: '*intelligence is the capacity of an agent to use computation, intended as the capacity to link perception to action in multiple possibly sophisticated ways, to increase biological fitness or to accomplish goals*. This view agrees with some authors (Gardner, 1992) claiming the existence of *multiple intelligences* involving different domains, e.g. verbal-linguistic, visual-spatial, logical-mathematical, naturalistic, and interpersonal intelligence. Despite the ideas of multiple intelligences are not supported by substantial experimental evidence (Visser, Ashton, and Vernon, 2006), domain-specific and domain-general cognition are generally accepted as distinct constructs, at least as the extremes of a continuum (Jacobs and Gärdenfors, 2017). This being said, here we focus on the Author's definition of (general-domain) intelligence as we think it might generate a useful research agenda for AI, which is worth following in parallel with other agendas investigating other domain-specific competences, e.g. the sensorimotor competence useful to control robots.

2. Our proposal: looking at principles of intelligence in the brain

Having broadly agreed on the target definition, we now focus on the research agenda deriving from it. Here our approach substantially diverges from the Author's one, pivoting on a logic-based reasoning approach ('*Non-Axiomatic Logic*') and taking distance from the brain. Indeed, we aim to

build AI systems with general-domain intelligence by looking at the *general principles that underlie intelligence in the brain*. The rationale of this is that the space of possible mechanisms that might implement general-domain intelligence is huge, and thus the viable solutions are extremely difficult to find (Baldassarre et al., 2017). A way to restrict the search space is thus to look at the brain, the only known system able to express general-domain intelligence, found by evolutionary selection in millions of years. In line with the Author, we qualify this statement by observing that we still poorly understand the complex detail processes of the brain, and so its study cannot suggest already-operationalised specific algorithms. However, we think that the analysis of the brain can indicate very useful *general principles*, to be formalised in detail, allowing a remarkable reduction of the space of solutions for building AI.

A first principle useful to design AI systems that we derive from the brain is that the *flexibility* of cognition required to cope with incomplete knowledge, is based on processes underlying what is called *goal-directed behaviour* (Balleine and Dickinson, 1998; Mannella, Gurney, and Baldassarre, 2013). Goal-directed behaviour aims to accomplish *goals* (desired future states of the world) and is based on processes that flexibly compose action-outcome chains (planning) to accomplish them. This composition is performed within the mind on the basis of general-purpose models of the world. Goal-directed behaviour is complementary to *habitual behaviour*, based on rigid stimulus-response associations. Our stress on goal-directness and planning agrees with the definitions of intelligence proposed by some fathers of AI highlighting the importance of 'ends' or 'goals' (Newell and Simon, 1975; McCarthy, 1988), reported but not expanded by the Author. In this respect, many studies show a high correlation between the flexibility of general-domain intelligence and a set of goal-directed processes called *executive functions*, e.g. involving inhibitory control, working memory, and cognitive flexibility (Diamond, 2013). The study of these processes might give important information to specify how (general-domain) intelligence might be accomplished.

A second principle is that goal-directed processes, although necessary, are not sufficient for flexibility. For example, classic AI planning implements goal-directed behaviour based on goals and world models, but the resulting systems are quite rigid. This is a limitation because ecological conditions, as also stressed by the Author, always involve challenges such as novel states, goals, and needed actions, on which the agent lacks knowledge (Santucci, Baldassarre, and Mirolli, 2016). We posit that a main way the brain uses to face this lack of knowledge is through *processes of manipulation of internal representations of knowledge* (alongside actively seeking such knowledge in the external environment, an important issue we cannot further expand here for lack of space, see (Baldassarre and Mirolli, 2013)). This manipulation allows the brain to perform different operations on internal representations, for example abstraction over details or selection of specific parts of an object, so as to form the needed new representations starting from the acquired ones (these operations might also be closely linked to conscious processing (Baldassarre and Granato, 2019)). This internal manipulation of representations, at the basis of *imagination* (Seepanomwan et al., 2015) and *problem solving*, allows the brain to modify the previously acquired knowledge to produce the lacking knowledge needed to accomplish novel goals or familiar goals in novel conditions/domains.

A third principle is that intelligent systems should be based on sub-symbolic representations and parallel distributed processing, as those of neural networks, rather than on symbolic representations and logic inference as proposed by the Author. The importance of this principle, having far roots in early AI (McCulloch and Pitts, 1943; McClelland, Rumelhart, and the PDP research group, 1986), is being demonstrated by the recent overwhelming success of deep neural networks (LeCun,

Bengio, and Hinton, 2015). Parallel distributed processing is central to our proposal as it allows the manipulation of internal representations by leveraging properties, such as *generalisation* and *generativity*, that are not possible with symbolic representations (Baldassarre, 2002; Graves and et al., 2016; LeCun, Bengio, and Hinton, 2015).

3. Flexibility as manipulation of internal representations: an example model

Our perspective also allows the specification of the concept of *adaptation*, a key element of the Author's definition of intelligence. The Author suggests that adaptation refers to ontogenetic changes (vs. evolutionary changes), involves changes of the environment and not only of the agent, and is possible only when the new challenges are similar to past ones. We think these features do not fully capture what is needed to cope with situations involving partial knowledge (incidentally, we prefer to talk of *flexibility* rather than 'adaptation', a term mainly referring to biological intelligence but less suitable for AI). Our proposal allows the specification of flexibility (adaptation) by stating that in order to flexibly solve goals for which it lacks knowledge, an intelligent agent not only searches information and knowledge in the environment but it also actively builds it internally (manipulation of internal representations).

In (Granato and Baldassarre, 2019), we propose a computational model that starts to illustrate our proposal. The model controls an artificial agent that pivots on a generative neural network (a Restricted Boltzmann Machine (Hinton, 2002)). The model is able to manipulate internal representations and self-generate input images corresponding to them (imagination). The model is tested with a simulated *Wisconsin Cards Sorting Test* (WCST (Heaton et al., 2000)) used in neuropsychology to measure *cognitive flexibility*. The test requires the agent to discover an unknown rule to match deck cards to one of four target cards, either by colour, form, or size, based on a positive/negative feedback received in repeated matching trials. The rule changes after some trials thus changing the condition to accomplish the goal of 'getting the positive feedback.' The change requires a flexible switch of the sorting rule. The model uses a mechanism akin to reinforcement learning to select the specific feature of the cards (internal manipulation) with which to generate modified images of the cards, reproducing only their colour, form, or size, used for matching. Although very simple, the model exemplifies how the manipulation of internal representations might boost flexibility in a goal-directed agent lacking knowledge to accomplish a given goal.

We conclude this commentary by observing that our proposal suggests a redefinition of the target concept as follows: '*General-domain intelligence is the capacity of goal-directed agents to flexibly accomplish novel goals in novel conditions/domains by building the knowledge they lack through the manipulation of internal representations and by actively seeking such knowledge in the external environment.*'

References

Baldassarre, G. and Granato, G. 2019. Representation Internal-Manipulation (RIM): A Neuro-Inspired Computational Theory of Consciousness.

Baldassarre, G. and Mirolli, M. 2013. Intrinsically Motivated Learning Systems: An Overview. In Baldassarre, G. and Mirolli, M., eds., *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. 1–14.

Baldassarre, G., Santucci, V. G., Cartoni, E., and Caligiore, D. 2017. The architecture challenge: Future artificial-intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction. *Behavioral and Brain Sciences* 40(40):e254.

Baldassarre, G. 2002. *Planning with neural networks and reinforcement learning*. PhD thesis, Computer Science Department, University of Essex, Colchester, UK.

Balleine, B. W. and Dickinson, A. 1998. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37(4-5):407–419.

Diamond, A. 2013. Executive functions. *Annual review of psychology* 64:135–168.

Gardner, H. 1992. *Multiple intelligences*, volume 5. Minnesota Center for Arts Education.

Granato, G. and Baldassarre, G. 2019. Goal-directed top-down control of perceptual representations: A computational model of the Wisconsin Card Sorting Test. In *Conference on Cognitive Computational Neuroscience*. 13-16 September 2019, Berlin Germany.

Graves, A. and et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538:471–476.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., and Curtiss. 2000. *WCST: Wisconsin card sorting test: manuale.* Firenze O.S., Italy.

Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation* 14(8):1771–1800.

Jacobs, I. and Gärdenfors, P. 2017. The false dichotomy of domain-specific versus domain-general cognition. *Behavioral and Brain Sciences* 40.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature* 521:436–444.

Mannella, F., Gurney, K., and Baldassarre, G. 2013. The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. *Frontiers in Behavioral Neuroscience* 7(135):E1–29.

McCarthy, J. 1988. Mathematical logic in artificial intelligence. *Daedalus* 297–311.

McClelland, J. L., Rumelhart, D. E., and the PDP research group. 1986. *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge, MA: The MIT Press.

McCulloch, W. S. and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5(4):115–133.

Newell, A. and Simon, H. A. 1975. Computer science as empirical inquiry: Symbols and search. *Philosophy of psychology* 407.

Santucci, V. G., Baldassarre, G., and Mirolli, M. 2016. GRAIL: A goal-discovering robotic architecture for intrinsically-motivated learning. *IEEE Transactions on Cognitive and Developmental Systems* 8(3):214–231.

Seepanomwan, K., Caligiore, D., Cangelosi, A., and Baldassarre, G. 2015. Generalization, decision making, and embodiment effects in mental rotation: a neurorobotic architecture tested with a humanoid robot. *Neural Networks* 72:31–47.

Visser, B. A., Ashton, M. C., and Vernon, P. A. 2006. g and the measurement of Multiple Intelligences: A response to Gardner. *Intelligence* 34(5):507–510.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 24-26, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-18
Accepted 2019-11-07

AI: A Crowd-Sourced Criterion. A Commentary on Pei Wang's Paper "On Defining Artificial Intelligence"

Istvan S. N. Berkeley ISTVAN@LOUISIANA.EDU
Philosophy
The University of Louisiana at Lafayette
Lafayette, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Since Wittgenstein's discussion (Wittgenstein, 1953) of the term 'game,' and other terms, in his *Philosophical Investigations*, there has been an air of suspicion over definitions amongst philosophers. The kinds of concerns which Wittgenstein raised seem to apply to Wang's proposed definition (Wang, 2019). While Wang's arguments in favor of developing a definition of artificial intelligence appear to be well-founded, his proposed definition is problematic in a number of ways. The chief defect is that it is too vague (see Pelletier and Berkeley (1995)). The definition Wang proposes is,

> "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources."

Although on the face of it, this sounds laudable, it becomes problematic when applied in practice. For instance, Weizenbaum's classic program ELIZA (Weizenbaum, 1966) appears to trivially satisfy this definition, in so much as, given that the system is nearly fifty years old, it clearly ran with limited resources, at least as compared to modern computational systems. However, as ELIZA had a means of handling a wide range of inputs, often by turning assertions into questions, it clearly had a means of 'adapting to its environment.' Yet, most contemporary theorists would be reticent to ascribe intelligence to ELIZA. So, the definition fails by including too many information-processing systems.

Another way that the definition fails is by also excluding certain information-processing systems that we might plausibly wish to count as intelligent. Consider the case of the system described by Ciresan et al. (2012). This system was designed to classify traffic signs. How well it would function trying to process traffic signs from other countries, or jurisdictions, is not known. However, if it failed to handle them in a sensible manner, then it would appear to have failed to 'adapt to its environment' and would thereby fall outside the class of intelligent artifacts. Indeed, given that many information-processing systems are built, or trained, to perform specific tasks, it appears quite likely that relatively few systems could satisfy Wang's definition of intelligence. This is surely a defect of the proposal.

Does this mean that the quest for a clear understanding of the phrase 'Artificial Intelligence' will always remain elusive? Perhaps not.

Famously, in his paper 'Computer Machinery and Intelligence,' Turing (1950, p. 442) remarked that,

> "...I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

Given that 'thinking' is a central component of intelligence, Turing's suggestion may offer a way forward to a better specification of the phrase 'Artificial Intelligence.'

Berkeley and Rice (2013) considered the extent to which Turing's proposal had become true. They used a variety of corpus linguistics methods to see whether people actually applied mental terms to computers. Although they did not look at the use of the term 'intelligence,' they did look for the actual use of terms like 'thinks,' 'believes,' and 'knows' in a variety of corpora. They also contrasted the use of these terms, with respect to computers, with the same terms used with respect to dogs. Their results seemed to show that, to some degree, Turing's suggestion had come true. There is no reason to believe that a similar study, looking at 'intelligence,' might not yield positive results, if not at the present time, then at some point in the future.

Now, one complicating factor which arises with the phrase 'artificial intelligence' with this proposed method is that it is a phrase that may be used for marketing purposes. The phrase 'artificial intelligence' has historically, from time to time, been subject to media and marketing enthusiasms (see Boden (2006)) for an extensive and detailed discussion). However, this crowd-sourced approach could at least avoid the issues outlined above.

By way of conclusion, this leaves the question of what is to be made of Wang's definitional proposal? A straightforward way of handling it, which is consistent with the considerations outlined here, is to consider it less as being definitional, and rather interpret it as being aspirational. It suggests a way that future putatively 'intelligent' computational systems could, or should be, assessed. A further advantage of the strategy suggested here is that it provides an empirical methodology of determining the extent to which people are prepared to adopt an 'intentional stance' toward computational systems (see Dennett (1971)).

References

Berkeley, I. S. N. and Rice, C. 2013. Machine Mentality? In Müller, V. C., ed., *Theory and Philosophy of Artificial Intelligence*. Berlin: SAPERE/Springer. 1–15.

Boden, M. 2006. *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford U. P.

Ciresan, D., Meier, U., Masci, J., and Schmidhuber, J. 2012. Multi-column deep neural network for traffic sign classification. *Neural Networks* 32:333–338.

Dennett, D. 1971. Intentional Systems. *Journal of Philosophy* 68:87–106.

Pelletier, F. J. and Berkeley, I. S. N. 1995. Vagueness. In Audi, R., ed., *The Cambridge Dictionary of Philosophy*. Cambridge: Cambridge, U.P. 825–827.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 236:433–460.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Weizenbaum, J. 1966. ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the Association for Computing Machinery* 9:36–45.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell's.

Journal of Artificial General Intelligence 11(2) 27-30, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-11-11
Accepted 2019-11-11

A Definition of Intelligence for the Real World?

François Chollet FCHOLLET@GOOGLE.COM
Google
Mountain View, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

"On defining artificial intelligence" by Pei Wang (2019) proposes a definition of the nature and goals of AI. Wang starts by arguing the incontestable benefits of explicitly defining AI—after all, any researcher who is actually trying to explain or create "intelligence" must necessarily hold a working definition of it, whether or not they attempt to make this definition explicit. Wang identifies specific criteria that a good working definition should match, using Carnap's approach for defining the murky concept of "probability" (Carnap, 1950): a good definition should map well to what people usually mean by "intelligence," it should draw a sharp boundary between what is intelligence and what is not, it should be fruitful to research, and it should be simple. Wang follows-up with a historical perspective on AI and an insightful taxonomy of different conceptualizations of intelligence in AI research (e.g. principle AI, structure AI, etc.), which was my favorite part of the paper.

Wang then offers his own definition: intelligence is a lifelong learning and adaptation process happening within an individual agent, driven by embodied experience, which occurs in real-time, under insufficient information and resources (such as computing resources). It is open-ended, may involve co-adaptation, and may not necessarily converge, as it takes place in an ever-changing, non-repeatable world. Finally, Wang discusses how his ongoing work on NARS and NAL (Durisek, 2014) is motivated and guided by this definition.

Coming from a cognitive developmental robotics background, I find myself highly sympathetic to Wang's view of cognition. While this vision of intelligence has not been very popular in the field of AI, it is historically ancient within the field of psychology. Wang mentions Jean Piaget, who built his own theory of intelligence development based on a similar vision in the 1940s (Piaget, 1947), but these ideas preexisted Piaget—we note that one of the very first researchers to ever attempt to rigorously characterize and measure intelligence, Alfred Binet, defined intelligence in rather similar terms (Binet and Simon, 1916, pp. 42–43):

> "It seems to us that in intelligence there is a fundamental faculty, the alteration or the lack of which, is of the utmost importance for practical life. This faculty is [...] the faculty of adapting one's self to circumstances."

However, it seems to me that Wang's view is more of a high-level vision than a precise and useful definition of intelligence. Overall, Wang's definition, while grounded in a very reasonable and even wise vision of intelligence, falls short of its own goals of "drawing a sharp boundary" and "being fruitful," due to insufficient formalism and excessive reliance on implicit semantics.

1. The boundaries of intelligence

As Wang points out, "drawing a sharp boundary" represents the ability to use the definition (and the definition alone) to categorize what is intelligence and what is not. As such, Wang uses his definition to argue that a "species" is not an intelligent system, but this argument relies on superficial semantics and excessive anthropocentrism (e.g. what defines an "individual" and "experience" in common language), rather than on any intrinsic property of his definition. The definition alone is not enough to draw this boundary. Taking a more formal approach, one can argue that a species may be modeled as a coherent system which undergoes adaptation driven by embodied experience—much like a human could also be modeled either as a coherent system or as a population of cells—which would fit the original definition. The less formal the definition, the more it will be open to interpretation, subjectivity, and conceptual interference from ungrounded semantics. If a concept is to be understood clearly, it should be fully described in absolutely precise terms: it should be made formal.

2. The fruitfulness of a definition

Wang notes rightly that the essential purpose of a definition is to be useful. But what would it mean, in practical terms, for a definition of intelligence to be useful? This is not explicitly explained by Wang. But let me advance my answer:

1. It should categorize and measure. The definition should be sufficiently precise to tell if a given system possesses, or not, a degree of intelligence. It should be sufficiently precise (in fact, quantitative) to be used to judge whether system A is more intelligent than system B. If there are different kinds of intelligences, it should feature a taxonomy and concrete methods for identifying the kind of intelligence possessed by a system (if any), and quantitatively comparing it to that of other systems.

2. It should guide. A good working definition should not merely describe (however precisely and accurately), it should be capable of serving as a North Star: a way to discard certain research avenues and to point to others, and to measure the progress being made at every step. It should make it clear whether subfield A is more or less likely to lead to intelligence than subfield B. It should highlight approaches that may not be of practical interest today but that show potential. It should be informative with regard to how far away we are from AGI.

3. It should explain. It should be illuminating with regard to why biological intelligent systems work the way they do, and ultimately, with regard to how to implement intelligence. A true definition must have explanatory power, and thereby should lead to progress not just in AI but also in neuropsychology. This idea is closely related to the "principle AI" brand of AI conceptualization.

Wang's definition cover 1, albeit not quite sufficiently precisely (e.g. it isn't formal or quantitative and its boundary relies on common sense English semantics), but it does not substantially attempt to do 2 or 3. As a result, it falls short of being fruitful.

3. From the generic to the specific

In addition, while Wang spends several pages discussing differences and shared ground with other diverging visions of intelligence, it is not clearly argued why Wang's definition should be preferred over any alternative definition that would fit the same vision. There could be multiple definitions that would be compatible with Wang's overall vision yet that could be more precise or more formal. In fact, despite almost entirely agreeing with Wang's overall high-level vision (intelligence is adaptation under high uncertainty, it is contextual, situated and embodied, it involves both the agent adapting itself to its environment and the agent adapting its environment, it must operate in real time, it is a lifelong learning process, it is open-ended and may not converge), I find myself mostly disagreeing with the architecture and specifics of NARS and NAL.

There is a large jump in Wang's argument between the vagueness of his working definition and the high specificity of his work on NARS and NAL—to such an extent that one may feel that the arguments behind Wang's definition were retrospectively conceived to justify the work on NARS and NAL. I would suggest drawing a clear intellectual path from the original definition to the concrete project, by iteratively listing at each "level of conceptual resolution" what possible formal choices or implementation can be made to make the abstract level more concrete, and clearly arguing why Wang's opinionated choices are better than alternative possibilities. And, perhaps, some of these other choices may prove interesting too.

4. Conclusion

In conclusion, despite the fact that the proposed definition falls short of its stated goals, I think Wang's vision is worthy of much more attention within the AI community. It is aligned with the work of people (such as Piaget) who had, decades ago, a much more grounded, nuanced, and deeper understanding of human cognition than many of the gradient-descent maximalists of recent days. Perhaps a way to achieve greater attention would be to propose more formal and more applied specialized definitions, and to offer concrete benchmarks or challenges to explore the practical consequences of working under these definitions in specific application settings, in the spirit of (Hernández-Orallo, 2017; Chollet, 2019). Connectionism, too, was shunned at one point, but it quickly became a darling once again the moment it found a useful domain of application, which was catalyzed by key benchmarks and challenges in 2011-2012 (in particular ILSVRC).

Ultimately, practical impact in the real world is the scale by which the value of a working definition of AI will be weighted.

References

Binet, A. and Simon, T. 1916. *The development of intelligence in children (The Binet-Simon Scale)*. Williams & Wilkins Co.

Carnap, R. 1950. *Logical foundations of probability*. Chicago: The University of Chicago Press.

Chollet, F. 2019. The Measure of Intelligence. arXiv:1911.01547 [cs.AI]. `https://arxiv.org/abs/1911.01547`.

Durisek, P. 2014. Pei Wang's Non-Axiomatic Reasoning System. `https://www.applied-nars.com/`.

Hernández-Orallo, J. 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press.

Piaget, J. 1947. *The psychology of intelligence*. Routledge & Kegan Paul Ltd.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 31-34, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-26
Accepted 2019-11-07

Defining Artificial Intelligence: Resilient Experts, Fragile Geniuses, and the Potential of Deep Reinforcement Learning

Matthew Crosby M.CROSBY@IMPERIAL.AC.UK
Leverhulme Centre for the Future of Intelligence
Imperial College London
London, UK

Henry Shevlin HFS35@CAM.AC.UK
Leverhulme Centre for the Future of Intelligence
University of Cambridge
Cambridge, UK

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Wang's definition of Artificial Intelligence is developed via careful and thorough abstractions from human intelligence. Motivated by the goal of building a definition that will be genuinely useful for AI researchers, Wang ultimately provides an agent-centric definition that focuses on systems operating with insufficient knowledge and resources. The definition captures many key components of intelligence, but we suggest that task success could play a slightly larger role. This brings the definition closer in line with our use of the term with animals and human experts, and also further aligns the definition's associated research framework with the subfield of deep reinforcement learning aimed at general intelligence.

1. Introduction

Wang (2019) proposes a working definition of artificial intelligence based on a system's ability to adapt to its environment under certain resource constraints (AIKR). The definition is accompanied by a useful exploration of different perspectives in AI that vary with respect to how they abstract from human intelligence. One *prima facie* challenge for this approach is the worry of anthropocentrism: the space of possible intelligences is vast, and anchoring the definition to humans risks blinding us to a large portion of it. However, to the credit of Wang's approach, humans are used as *exemplars* of the explicandum, not necessarily its sole arbiters.

We believe that Wang's definition picks out key elements of intelligence. The insufficiency assumption allows for a Principle-AI-based definition that does not fall into the trappings of extreme versions, such as AIXI (Hutter, 2005), that revolve almost exclusively around task solving. Its move away from purely capability-based definitions is also positive. However, we suggest that some reference to capabilities would still be of benefit to Wang's approach. We first suggest that this will bring it closer to our usage with animals and experts, then look at how an updated definition aligns with the intuitions of AI researchers working towards general intelligence in deep reinforcement learning.

2. Intelligence and insufficiency

We agree with Wang that the ability to deal efficiently with scenarios in which knowledge and resources are lacking is a key marker of intelligence. One fact that this definition captures elegantly is that intelligence is not just a matter of completing tasks: if one's goals are simple and the environment stable, it is possible to thrive via relatively straightforward strategies, or a 'Resilient Idiot' approach, as exemplified by organisms like nematode worms or sessile shellfish. However, the definition is arguably too restrictive as currently stated. In particular, we suggest it risks leaving out two types of intelligent system that we term 'Resilient Experts' and 'Fragile Geniuses'. We define a Resilient Expert as a system that has rich stores of knowledge and multiple redundant mechanisms for solving any problems it encounters. Much like the Resilient Idiot, the Resilient Expert simply does not encounter insufficiency or uncertainty. Unlike the case of the Resilient Idiot, however, this resilience is a hard-won achievement for the Resilient Expert, and is founded upon expensive investments in knowledge and resources.

A wide range of biological organisms that we are inclined to describe as intelligent might plausibly qualify as Resilient Experts. As a simple example, note the extremely robust navigational capacities of animals such as bees. Bees make use of environmental landmarks, track the location of the sun, calculate the polarity of light (useful on overcast days), and track the speed and vector of prior movement using dedicated neural assemblies (Gould and Gould, 1988; Stone et al., 2017). The resilience and complexity of the bee's complex navigational toolkit bespeaks a sophisticated and intelligent biological agent. This is in spite of the fact that the bee (at least qua navigation) rarely if ever has insufficient resources to carry out its tasks. We recognise that the case of the bee just provided involves a single task, namely navigation, and Wang notes that AIKR applies to "the overall situation, not on every task, as there are surely simple tasks for which the system's knowledge and resources are relatively sufficient." However, we think it reasonable to imagine that there could be Resilient Experts whose knowledge and resources were bountiful in every domain yet still qualified as intelligent.

A second form of intelligent system that Wang's definition of intelligence might not easily cover in its current form is the Fragile Genius. By a Fragile Genius, we mean a system that struggles with uncertainty and insufficiency, but which (intuitively) constitutes an instance of intelligence by virtue of specialising towards some particularly impressive or complex goal. Consider a brilliant but eccentric composer who writes symphonies of dazzling beauty, creativity, and complexity, but who is incapable of reliably feeding or clothing themselves or even obtaining materials for producing their compositions. They are wholly dependent on the cooperation of the external environment for their continued thriving and do not adapt well under AIKR conditions.

Most of us are Fragile Geniuses. Our way of life depends on rich cultural and technological knowledge and complex co-ordination and specialisation of roles. Without the scaffolding of our cultural knowledge and technology most of us would struggle to adapt to even basic tasks like obtaining food, constructing shelter, or treating injuries (Henrich, 2017). Nonetheless, it is surely false to suggest that the fragility of modern life is such that fewer demands are placed on our intelligence. Rather, the acquisition of rich cultural storehouses of knowledge and specialisation of individuals has enabled us to develop skills and proficiencies unthinkable for our neolithic ancestors, including such elevated achievements as quantum mechanics, aeronautical engineering, and the Baked Alaska.

While adaptation under AIKR is highly indicative of intelligence, some very intelligent agents—the Resilient Experts—have managed to avoid uncertainty and insufficiency all together via complex redundant systems, while others—the Fragile Geniuses—struggle in the face of these factors. Hence we would suggest that task complexity and task solving ability might be given some more prominent role in the definition, even if uncertainty and insufficiency remain the unifying theme.

3. Deep Reinforcement Learning and Artificial Intelligence

Measuring progress towards intelligence is hard, so AI research tends instead towards measurable tasks with determinate success conditions. This leaves current mainstream research strands at odds with the kind of definition Wang is proposing. However, with task capability and complexity even a small part of the picture, we believe that certain strands of current deep reinforcement learning research do qualify as working towards intelligence. The goal *is* learning that is lifelong, cumulative, open-ended, and multi-objective; it's just a long way away.

Wang suggests that the optimality of many machine learning algorithms goes against AIKR. Whilst many Deep Reinforcement Learning (DRL) algorithms are based on convergence proofs, Deep Learning usually involves non-linear approximations and DRL is often applied in situations where assumptions required for the proofs do not hold. It is commonly assumed that environments are fully observable Markov Decision Processes, but in practice this is rarely the case (Arulkumaran et al., 2017).

Lifelong and continual learning is a growing area of research in the DRL community, starting with methods to prevent catastrophic forgetting (Kirkpatrick et al., 2017), where neural networks will sometimes jump away from a favourable weight space and 'forget' everything they have previously learned. Whilst the environments used for research often do not deviate too far from standard machine learning (Lopez-Paz and Ranzato, 2017), progress is being made towards the introduction of new continual learning paradigms (Khetarpal et al., 2018). There are many open issues, and research is still in its infancy (Schaul et al., 2018), but first steps are being taken towards testing systems in finite, open, and real-time settings (Beyret et al., 2019). Many 'intelligent' AI researchers are working with a similar definition of artificial intelligence. It is perhaps ironic that their goal is frustrated by insufficient knowledge and resources.

4. Conclusion

Wang's definition picks out important components of intelligence and sets an interesting research agenda. It scores well on the criteria of similarity, exactness, fruitfulness, and simplicity, but could make stronger requirements on task capabilities. Doing so brings it closer to our usage for resilient experts, fragile geniuses, and even many DRL researchers.

Acknowledgments

This work was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067.

References

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34(6):26–38.

Beyret, B., Hernández-Orallo, J., Cheke, L., Halina, M., Shanahan, M., and Crosby, M. 2019. The Animal-AI Environment: Training and Testing Animal-Like Artificial Cognition. arXiv:1909.07483 [cs.LG]. https://arxiv.org/abs/1909.07483.

Gould, J. L. and Gould, C. G. 1988. *The Honey Bee.* Scientific American Library.

Henrich, J. 2017. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hutter, M. 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.

Khetarpal, K., Sodhani, S., Chandar, S., and Precup, D. 2018. Environments for Lifelong Reinforcement Learning. arXiv:1811.10732 [cs.AI]. https://arxiv.org/abs/1811.10732.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114(13):3521–3526.

Lopez-Paz, D. and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 6467–6476.

Schaul, T., van Hasselt, H., Modayil, J., White, M., White, A., Bacon, P., Harb, J., Mourad, S., Bellemare, M., and Precup, D. 2018. The Barbados 2018 List of Open Issues in Continual Learning. arXiv:1811.07004 [cs.AI]. https://arxiv.org/abs/1811.07004.

Stone, T., Webb, B., Adden, A., Weddig, N. B., Honkanen, A., Templin, R., Wcislo, W., Scimeca, L., Warrant, E., and Heinze, S. 2017. An anatomically constrained model for path integration in the bee brain. *Current Biology* 27(20):3069–3085.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 35-40, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-11-01
Accepted 2020-01-16

Towards a Canonical Theory of General Intelligence

John Fox JOHN.FOX@ENG.OX.AC.UK
Lincoln College, Oxford University
Oxford, UK

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

In his 1990 book, *Unified Theories of Cognition*[1] AI founder Allen Newell attempted to bring together ideas from psychology, AI and computer science into a single framework with the twin aims of understanding human intelligence and building general intelligent systems. Professor Wang and I share Allen's vision, but in this short commentary I will try to bring out some of the differences as well as similarities in our views. In particular I will suggest that the cognitive sciences are converging on a view of intelligence that I call *Canonical General Intelligence* that may help to articulate the notion of intelligence in a way that informs our understanding of human intelligence and provides practical foundations for developing flexible, perhaps even general AI systems.

1. Intelligence and cognitive science

Coming up with universally agreed definitions of terms like intelligence, mind, rationality, etc. has been famously difficult for traditional disciplines like psychology and philosophy. Nowadays it is further complicated by the fragmentation of the "cognitive sciences," which originally sought to unify psychology and AI with philosophy, neuroscience, linguistics, etc. but which are now composed of many communities whose interactions are haphazard at best and often rivalrous.

"Cognitive science is the interdisciplinary scientific study of the mind and its processes."[2] It investigates many aspects of human mental processes like reasoning, problem-solving, decision-making, and planning, which most of us would agree are exemplars of intelligent action whether natural or artificial. Many cognitive scientists also expect a unified theory to cover learning, perception, natural language and processes that permit an agent to cope with the changing, unpredictable, complex and critical circumstances that humans and other animals, chatbots, robots and increasingly autonomous systems face in real-world environments.[3]

Prof. Wang's discussion (Wang, 2019) covers human intelligence (in education, psychology), artificial intelligence (design, engineering) and abstract principles of rational minds (philosophy in a very broad sense). I have tried to be similarly eclectic but with the difference that I have drawn on lessons learned in a real-world domain that raises challenging questions for all these fields. The domain is medicine, one of the largest and most complex fields that humans work in.

1. https://en.wikipedia.org/wiki/Unified_Theories_of_Cognition accessed September 2019.
2. https://en.wikipedia.org/wiki/Cognitive_science accessed October 2019.
3. For many cognitive scientists sensory, motor and even affective functions are also in scope even though their "cognitive" status is debatable.

I agree with Prof. Wang that "when working on a model of the mind, it will be nice if some results can find practical applications; when the direct goal is to solve a real-life problem" (Wang, 2019, p. 14). I would in fact put it more strongly; medicine is an exciting domain in which to do cognitive science research, a good place to formulate hypotheses and test theories. Medicine draws on a vast diversity of knowledge and human skills and requires many different forms of intelligence.

2. Let a hundred flowers bloom

The aspiration to achieve a general framework for understanding and building intelligent systems has been especially difficult because there are so many different schools of thought about the nature of mind and how it should be studied. The "flower" picture below illustrates some of the most prominent research traditions for which an attempt to understand the nature of intelligence is central. This differentiation and proliferation of schools has often produced profound differences in theoretical assumptions, explanatory paradigms and research methods which significantly impede interdisciplinary communication and progress towards a general science of intelligence.

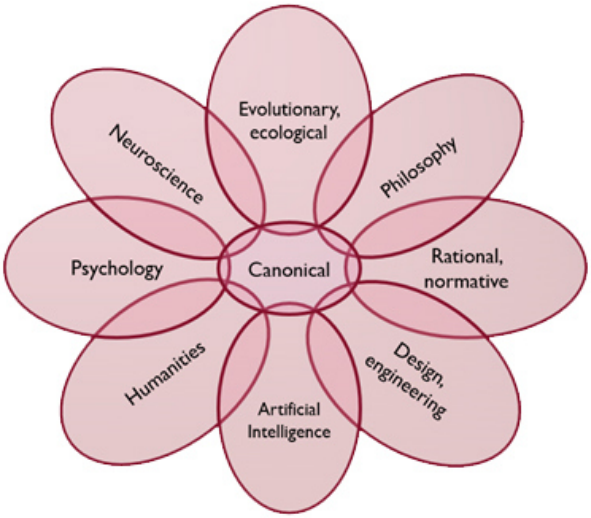

Figure 1: Some distinctive viewpoints on the nature of mind and, by extension, intelligence.

3. Common ground, Canonical theories

I have a longstanding interest in articulating a common ground theory that can provide a bridge between such different perspectives. This is illustrated schematically by the intersection of the flower petals which I have labelled "canonical" (Fox, Beveridge, and Glasspool, 2003; Fox, Cooper, and Glasspool, 2013). Whether a canonical theory of intelligence actually exists is open to question but Wang has summarised just the kind of general capabilities that many of us have in mind when we speak about AGIs:

> *"Such a system can be described as being driven by some tasks or problems . . . and can carry out or solve them . . . by taking some actions. The internal relations between*

> *the tasks and actions can be called the system's knowledge or beliefs. The system's information-processing activities are basically to choose and carry out proper actions to accomplish the existing tasks, and these activities cost computational resources, mainly the time for computation and action and the storage space for tasks and knowledge"* (Wang, 2019, p. 17).

This is a pretty good if rather informal description of what expert doctors (indeed all of us) do routinely and this sort of perspective has proved productive in thinking about the construction of practical AI systems that can carry out clinical tasks as well as or better than human professionals. Wang's description is a straightforward characterisation of everyday intelligence and would be common ground for many psychologists, neuroscientists, AI researchers and philosophers who ground their ideas in everyday concepts, like goals, knowledge and beliefs, and clinicians who are interested in such things. However, it is surely not precise enough to build a rigorous definition on and certainly not a general theory of intelligence.

Wang thinks, and most of us would agree, that we need to do more than just offer a scenario that exemplifies the phenomenon of intelligence. He suggests four dimensions that we should use to assess any definition: does it agree with our intuitions about what our intelligence is? ("Similarity to the explicandum"); does it unambiguously refer to behaviours that we agree are intelligent? ("Exactness"), is it productive in theoretical and/or practical terms? ("Fruitfulness") and is it clear and conceptually economical? ("Simplicity").

These criteria also seem to be good for judging whether a *theory of intelligence* is a good theory: a general theory should be intuitive (whatever your discipline), unambiguous, fruitful and parsimonious. For me these are desiderata for a theory that cuts across the cognitive sciences. To date a single common ground theory of (artificial) intelligence hasn't emerged; in Prof. Wang's view however the space of theories of intelligence that currently exist can be described in terms of a small number of orthogonal theoretical paradigms.

- *Behaviour-AI* concerns itself with the degree of similarity between the behaviour of a human agent and an artificial agent in similar circumstances.

- *Capability-AI* refers to "the intelligence of a system [in terms of] a set of problems it can solve."

- *Structure-AI* takes a "static" view of an intelligent system, as in a modular information processing architecture or even the anatomy of the brain.

- *Function-AI* focuses on the kinds of things that an agent, whether human or artificial, needs to be able to do when carrying out challenging tasks, like medical tasks (e.g. perceiving, reasoning, predicting, decision-making, planning, designing, acting, communicating, learning).

- *Principle-AI* is concerned with normative principles that AI designers should use to ensure the best possible outcome on practical tasks (e.g. classical or non-classical logics, probability axioms for reasoning under uncertainty, "rational" decision theory, or even meta-theories such as category theory).

The human cognitive sciences also seem to have so far evolved a number of general paradigms for studying, modelling and explaining the processes that underpin natural cognition, including *static*, *dynamic*, *epistemic* and *pathic*[4] paradigms.

- *Static information processing architectures* (typically visualised as box and arrow diagrams) have traditionally informed models of human cognition.
- *Dynamics*: a paradigm of cognitive modelling has emerged under the influence of AI in which *executable programs* are used to explain cognitive processes ("the program is the theory").
- *Epistemics*: Research on organisation of human memory and concepts has had fruitful interactions with AI research on *representation and use of knowledge*.
- *Pathics*: ideas like beliefs, desires and intentions from philosophy have an increasing role in understanding *mental states* and "folk psychology."

In our experience none of these theoretical paradigms is by itself anywhere near sufficient to capture the diversity of medical expertise or to design AI systems that can carry out the vast range of tasks that are routine for humans. A key question for me has been whether we can we develop a canonical theory of intelligence by observing, modelling and explaining human expertise, in the sense that it unifies the different paradigms while also satisfying Wang's criteria of *intuitiveness*, *exactness*, *parsimony* and *fruitfulness*.

4. Marr signatures

Prof. Wang rightly observes that we need to be more formal in saying what we mean by intelligence and articulating our theories (Wang, 2019, p. 5). The problem is finding a formalism that does the job.

He and I have independently identified a candidate for this purpose which we attribute to the late mathematician-turned-computational neuroscientist David Marr.[5] He is particularly remembered for his work on human and machine vision, but his general view of computational theory has been widely influential. This is expressed in terms of three complementary levels of analysis:

- what functions does the system perform? (e.g. what problem does it solve and why?)
- what algorithms does the system employ?
- how is the system physically realised? (e.g. neural tissue, silicon or quantum states)

In our own experience observing professional behaviour in clinical practice and building systems that have the capability and knowledge to carry out complex clinical tasks (*"AI-Complete Tasks"* (Wang, 2019, p. 12)) has benefited greatly from this way of decomposing the issues. It has in fact led to a way of using such signatures to characterise medical expertise in a way which seems to me to be intuitive and exact and has provided a practical basis for AI design (Fox, Beveridge, and Glasspool, 2003; Fox, Glasspool, and Modgil, 2006; Fox, 2017).

4. This is a neologism derived from *empathy* and *anthropic*.
5. `https://en.wikipedia.org/wiki/David\_Marr\_(neuroscientist)` accessed October 2019.

Wang's take on Marr is more focused on AI though in a similar spirit: *"to solve a problem by computation means we must 1. define the problem as a mapping from a domain of valid input values to a range of possible output values; 2. find an algorithm that carries out this mapping step by step, starting from the given input and ending with the corresponding output; 3. implement the algorithm in a computer system so as to use it to solve each instance of the problem"* (Wang, 2019, p. 16).

I believe that we can characterise many of the cognitive processes that underpin clinical expertise in terms of a surprisingly small set of general functions or "canons" that cut across theories, tasks and knowledge domains. The set of signatures in Figure 2 is an abstraction from the specialist knowledge of medicine and uses a general vocabulary that we are all familiar with (beliefs, goals, decisions, reasons, plans, arguments, actions, etc.). The terms used in the signatures can be formalised and given an exact semantics (e.g. (Fox and Das, 2000)) in multiple ways and instantiated with many different algorithms (e.g. logical, probabilistic, procedural), which can be implemented in many ways (e.g. biological or physical). The core elements of the canons are intelligible whether we are psychologists, neuroscientists or philosophers, computer scientists or designers of autonomous agents, through to journalists, novelists and the rest of us.

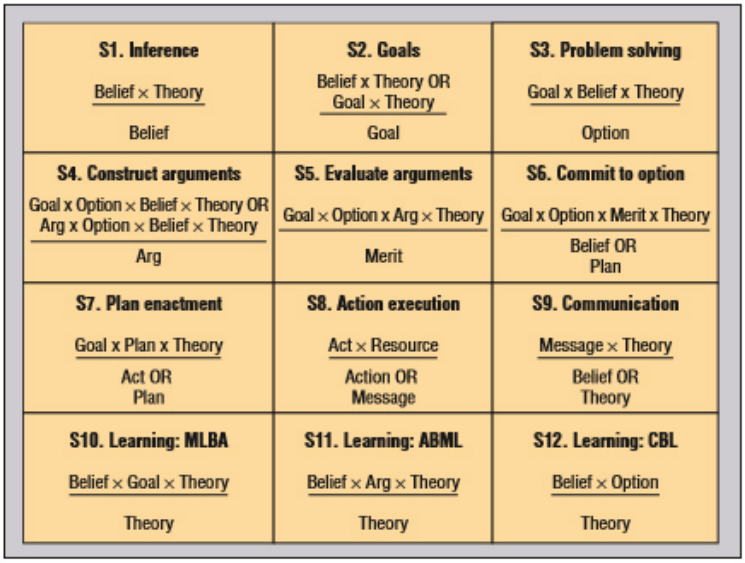

Figure 2: Canons of cognition: a set of generic cognitive functions that collectively form a theory of mental processes involved in decision-making under uncertainty and complex medical expertise (Fox, Glasspool, and Modgil, 2006) and have provided foundations for a general technology for building agents which are capable of operating above human expert level on a wide range of medical tasks (Fox, 2017).

On reading Prof Wang's discussion my first thought was that this set of signatures is an example of *function-AI* in his terms. However it can also be given a structural interpretation in terms of an agent architecture (Das et al., 1997; Fox and Das, 2000) and even neuro-anatomy (Shallice and Cooper, 2011, chapter 9). Furthermore this collection of signatures also provides a formal

foundation (Fox and Das, 2000, part 3) for an agent implementation language (Sutton and Fox, 2003) that has proved to be able to emulate human behaviour on a wide range of medical tasks.

My proposal about "Marr signatures" is not that we can capture a particular repertoire of intelligent capabilities with a single theory but that Marr's schema can express what is common ground for many research communities (what must an intelligent system do and why?) and do this with a clear, declarative notation (Figure 2) without committing to any a discipline-specific interpretation of the signatures (how the signature is refined as a mathematical algorithm or physical implementation).

We do not know whether this approach to formalising theories of intelligence in a canonical form will be useful beyond medical AI, though it appears to be usable for many other practical domains. The key point is that our multidisciplinary, diverse and sometimes fractious community needs some way of establishing common ground on which to have fruitful interdisciplinary discussions, and that Marr signatures may be a useful tool for this purpose.

References

Das, S. K., Fox, J., Elsdon, D., and Hammond, P. 1997. A flexible architecture for autonomous agents. *Journal of Experimental & Theoretical Artificial Intelligence* 9(4):407–440.

Fox, J. and Das, S. 2000. *Safe and Sound: AI in Hazardous Applications*. Menlo Park, CA: AAAI Press/MIT Press.

Fox, J., Beveridge, M., and Glasspool, D. 2003. Understanding Intelligent Agents: Analysis and Synthesis. *AI Communications* 16(3):139–152.

Fox, J., Cooper, R., and Glasspool, D. 2013. A canonical theory of dynamic decision-making. *Frontiers in Psychology* 4:150.

Fox, J., Glasspool, D., and Modgil, S. 2006. A Canonical Agent Model for Healthcare Applications. *IEEE Intelligent Systems* 21(6):21–28.

Fox, J. 2017. Cognitive systems at the point of care: The CREDO program. *Journal of Biomedical Informatics* 68:83–95.

Shallice, T. and Cooper, R. 2011. *The organisation of Mind*. Oxford University Press.

Sutton, D. and Fox, J. 2003. The Syntax and Semantics of the PROforma Guideline Modeling Language. *Journal of the American Medical Informatics Association* 10(5):433–443.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 41-44, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-18
Accepted 2019-11-05

Intelligence, Knowledge & Human-like Intelligence

John E. Laird LAIRD@UMICH.EDU
University of Michigan
Ann Arbor, Michigan, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

1. Defining Intelligence

I draw from Newell (1990, pp. 88–95), Legg and Hutter (2007), and others to define *intelligence*. I equate intelligence with *rationality*, where an agent uses its available knowledge to select the best action(s) to achieve its goal(s) within an environment. In this definition, intelligence is a measure of the optimality of behavior (actions) *relative* to an agent's available knowledge and its tasks, where a task consists of goals embedded in an environment. Intelligence can be measured for a single task, as is often standard practice in AI, but also for collections or sequences of tasks, where earlier tasks can provide knowledge that influences later task performance.

My motivation (consistent with Newell and others) is to define intelligence so that it is useful for evaluating any agent, relative to the tasks it pursues. This includes single-task agents, such as Schaeffer's Chinook, which has perfect intelligence for checkers, as well as humans, who can pursue many tasks over a lifetime, but usually have imperfect intelligence because of the difficulty of extracting and bringing to bear all their knowledge that accumulates over a lifetime. Importantly, intelligence is not a direct measure of specific internal processes and representations, although it indirectly measures their ability to bring knowledge to bear to control behavior.

This definition differs from concepts such as *adaptive intelligence*, *general intelligence*, or *human-level intelligence*. I see this as a good thing. Too often, the singular use of "intelligence" is overloaded so that it implicitly applies to either large sets of tasks or to especially challenging tasks (ones that "demand intelligence"), limiting its usefulness for more mundane, but still important situations. I propose that such concepts be defined using explicit modifiers to "intelligence." These modifiers can be associated with appropriate task properties, specifying collections and/or sequences of tasks. Legg and Hutter (2007) use this convention to define *Universal Intelligence* to be where intelligence is evaluated across all possible tasks. One can imagine using the same approach to define *Atari Game Intelligence*, *Autonomous Driving Intelligence*, or even *Creative Intelligence* for tasks that depend on using existing knowledge to create new knowledge.

Wang (2019) takes a different approach to defining intelligence, attempting to capture the essence of intelligence in properties of the processing and capabilities of an agent: an information processing system that adapts in the face of its own insufficient knowledge and resources. I assume that underlying Wang's definition is a desire to reserve the term "intelligence" for challenging processing, avoiding exhaustive searches or table lookups, and focusing on cases where knowledge must be discovered or transformed in complex ways to solve difficult problems. He further restricts

his definition to agents that are open to new tasks, which suggests he is thinking more about general intelligence than intelligence in general. I agree that the challenge of creating such agents is exciting, but casting these properties as definitional limits our ability to evaluate intelligence on individual tasks and simpler, but important AI systems. I'm not sure if this is his intention, but it appears that under his definition, non-learning systems, such as Chinook, Deep Blue, and Watson are not intelligent. Does he not include them in the study of AI? Although I disagree with his definition, I agree that research on agents with the properties he describes is important to the future of AI. It just isn't all there is. As suggested above, using my proposed definition, modifiers can capture such restrictions, possibly using "adaptive," or "creative" with the concurrent identification of sequences of tasks where the desired capabilities are necessary for high/successful performance.

2. Human-like Intelligence

My definition of intelligence can feel unsatisfying. On the surface it doesn't seem to provide direction for agent design beyond that agents should be rational. However, I contend that the story changes when we consider something approximating *human-like* intelligence. I somewhat crudely define "human-like" to refer to agents whose tasks and environment approximates humans: there are many different tasks that arise off and on over an extended lifetime, under varying temporal and computational resource constraints, in a complex dynamic environment that has exploitable regularities and is populated with other agents. In such agents, available knowledge is no longer just innate, pre-programmed knowledge, but includes the knowledge gained from experiences during task performance. More broadly, available knowledge also includes what the agent can learn from explicit exploration, as well as from interaction with other agents. Thus, for an agent to have high human-like intelligence, it must use all of these sources of knowledge. Below I explore the implications of these sources of knowledge on the design of highly intelligent human-like agents.

Innate Knowledge: In the simplest case, intelligence is a measure of how well a non-learning agent uses its innate knowledge to perform a task. Even in this case, the incompleteness theorem tells us that there are combinations of knowledge and tasks for which perfect intelligence is impossible. Furthermore, when there are limited computational resources, it can be challenging to access all relevant knowledge during task execution, especially from large knowledge bases. Much of early AI research explored how to represent, access, and process different types of knowledge so that an agent can use its relevant knowledge (and maximize intelligence) under resource bounds.

Task Experience: An agent's available knowledge increases with each interaction with its environment. Exploiting that knowledge involves extracting task relevant regularities so that they are available for future reasoning. If the space of tasks is known, an agent can be designed to extract only relevant regularities. However, if the space of tasks is unknown, as in human-like intelligence, an agent must learn not only new tasks, but also new task-relevant regularities. During task performance, processing constraints can make it impossible to perform all the analyses necessary to extract and encode the relevant knowledge. However, in the future time, when there is no time pressure, the agent can *retrospectively* analyze a trace of its behavior (if it has retained it!) given its existing knowledge, and extract regularities, even ones that were not known to be relevant during the original experience. More generally, a highly intelligent human-like agent would employ many of the processes that Wang identifies to determine the entailments of the knowledge it acquires.

Exploration: Beyond the agent's innate knowledge and experiences, there is the knowledge that an agent can gain through exploration with its environment. Thus, a highly intelligent agent will not

just passively learn as it performs its tasks, but will actively seek out knowledge in its environment that can aid future performance. Reinforcement learning incorporates exploration, but usually only in the service of repeated attempts at a task or sets of related tasks. A highly intelligent agent will expand exploration so that during any free time, it will deliberately engage in environmental interactions to increase its knowledge, proactively extracting knowledge not just for its current task, but for use in possible future tasks.

Personal exploration of a world as rich and large as ours exposes an agent to only a minuscule amount of the knowledge embedded in that world. For humans, language provides a means of accessing the experiences and knowledge of others, greatly increasing our intelligence beyond all other animals, even primates, who are unable to access such knowledge. Our intelligence is even further enhanced because we have access to books and other media, which provide efficient means of access to huge bodies of knowledge. Furthermore, our society not only accumulates knowledge, it also spends resources on deliberately creating and disseminating knowledge through research and educational institutions, making all of us more intelligent.

3. Discussion

My goal in this paper is to provide support for a common, straightforward definition of intelligence based on rationality, as well as to explore the implications of this definition for human-like agents. At one extreme, this definition makes it possible to compare the intelligence of different AI programs working on single tasks, such as puzzle solving or image identification, assuming they have access to the same knowledge. However, it is constrained in that intelligence is relative to a task and an agent's knowledge, so it is not meaningful to say that an agent that is an exceptional Go player is more or less intelligent than a mediocre checkers player. Nor is it meaningful to compare the intelligence of an agent that trained for days on a task to one trained for an hour, although it would be meaningful to compare them both after an hour of training.

At the other extreme, the definition allows meaningful, although imperfect, comparisons between humans at similar ages, and even between humans and non-humans. In both cases, the agents that better exploit their innate knowledge, their experiences, their environments, and the knowledge available from other agents are expected to display higher intelligence. Often we try to separate knowledge (or expertise and skill) from intelligence, but under this analysis, intelligence is highly dependent on the ability to discover, extract, and exploit knowledge.

A final, but intriguing question is whether highly intelligent human-like agents share an underlying cognitive architecture. As a step toward exploring that question, a group of us are developing an abstract theory of human-like minds, originally called the Standard Model of the Mind, but renamed the Common Model of Cognition (Laird, Lebiere, and Rosenbloom, 2017).

Acknowledgments

This work was supported by the AFOSR under Grant Number FA9550-18-1-0168. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of AFOSR or the U.S. Government.

References

Laird, J. E., Lebiere, C., and Rosenbloom, P. S. 2017. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*.

Legg, S. and Hutter, M. 2007. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines* 17:391–444.

Newell, A. 1990. *Unified Theories of Cognition*. Harvard University Press.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 45-46, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-08-29
Accepted 2019-11-05

A Review of "On Defining Artificial Intelligence"

Shane Legg SHANE@DEEPMIND.COM
DeepMind
London, UK

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

This (Wang, 2019) is a nice survey of views on the nature of machine intelligence and makes many good points. Let me focus here in my review just on those parts of the paper where my view diverges with what has been presented, or where I feel that I have additional comments that I would like to add.

The paper talks about the difference between AI and AGI. I think this is an important aspect of this discussion. AI has come to be a rather large and diverse field. That's possibly a rather good thing for an active and lively community and in this sense a lack of a crisp and agreed definition of AI doesn't have to be viewed as a problem. For example, economics (or many other large fields that you might care to name) might not have a single agreed upon definition, but that hardly matters. People can still come together and fruitfully develop new ideas and models, without needing to have a sense that they are all working towards some common goal or worrying that the borders of their discipline aren't neatly defined. The fact that economics bleeds over into aspects of governance, finance, politics and so on is part of the richness of the field. It is more when we are concerned with "building a general AI" (or AGI) that having a common and well defined goal becomes important for many of the reasons that this paper explains.

At the end of section 2.2.5 people interested in a principled approach to defining intelligence are referred to by some as having "physics envy." To repeat this claim is uncharitable in my opinion. One can be interested in clean and principled approaches to topics without suffering from any kind of envy, and to suggest otherwise has a whiff of ad hominem to it.

Section 3.1 references Hutter's 2005 definition of "universal intelligence." This should be Legg and Hutter (2007) as that is where "universal intelligence" appears.

In section 3.2, and also more in section 4.1, it is claimed that with the model of Legg and Hutter "whatever the agent does, the actions can only change the rewards it gets, but cannot change the environment." This is incorrect, or at least very misleading. In the Legg-Hutter model the "environment" can be any computable distribution over entire agent-environment interaction histories. Because this is a function over entire interaction histories it is equivalent to the environment being stateful. Thus you could have an environment where an agent has to stack some blocks in order to build a tower. You would normally say that this agent was indeed changing its environment. Indeed, the "environment" could even be a universal Turing machine that first reads a new program from the agent and then executes it, allowing the agent to entirely reprogram its environment from scratch! If anything, universal intelligence considers cases where agents can modify their environments in radical ways.

Finally, I'd like to comment on the idea that the notion of "insufficient knowledge and resources" should be part of the definition of intelligence. Clearly any real system will have resource and knowledge limitations. In which case, why do we need to make this a part of the definition? Simply create your system and let's see how capable it is! By bringing in this additional aspect we are mixing together what a system is capable of doing, with how it goes about achieving this. We don't do this in other domains: we don't say that speed is about how fast something moves through space given only finite energy. Or that a company's profits are about how much money it makes given limited resources. Yes, things in reality are always limited, but we don't build this into all our definitions. Dealing with this fact is sufficiently implied by the nature of reality.

When I have made this point in the past, some people then ask why I am interested in things like AIXI (Hutter, 2005). I see AIXI in the same way as I view Turing machines: as an abstract model which allows a certain kind of theoretical analysis, not as a blueprint for actually building a real system.

References

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer.

Legg, S. and Hutter, M. 2007. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines* 17:391–444.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 47-49, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-01
Accepted 2019-11-07

Intelligence and Agency

Peter Lindes PLINDES@UMICH.EDU
University of Michigan
Computer Science and Engineering
Ann Arbor, Michigan, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

1. Introduction

A clear working definition of anything must relate to a well-defined referent. Definitions of *artificial intelligence* tend to be confusing when they fail to distinguish between two common referents of this phrase. The first common usage, which we will call AI1, refers to the quality of intelligence in some man-made system. The second common usage, AI2, refers to the field of study which addresses systems of the AI1 sort. Thus a definition of AI2 depends on defining AI1, and a definition of AI1 depends on how we define *intelligence*. This commentary will focus on defining intelligence, and then how it relates to AI1 and AI2.

2. Wang's definition of intelligence

Pei Wang, in his paper entitled "On Defining Artificial Intelligence" (Wang, 2019), gives the following proposed definition:

> *Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources.*

This, of course, is a definition of *intelligence* in general, that could apply to humans, animals, or man-made systems, not a definition of AI *per se*. Since the paper talks about a wide range of things that could be intelligent, it is curious that this definition is centered on "an information-processing system." Technically, this term can be considered general enough to cover the whole range of systems Wang discusses, but its common usage tends to imply a computer system. This contradicts Wang's first requirement for a definition, that it have "similarity to the explicandum." It would be better to use a term without computer science implications, such as "agent."

Wang's definition than talks about a system's "capacity ... to adapt to its environment." Certainly a capacity to adapt is an important part of intelligence, but it is not the only important part. Before adapting, it seems an agent would need to act on a moment-to-moment basis in its environment. Perhaps "while operating" suggests ongoing action, but it seems a weak way to say it. Adaptation should not be just to the environment, but also to the agent's own internal needs and goals, which can change over time, as in the development of a child.

Wang says the system must operate "with insufficient knowledge and resources." Later in the paper he expands on this concept, calling it "the *Assumption of Insufficient Knowledge and*

Resources (AIKR)." What does it mean that they are "insufficient?" Certainly any finite agent that must operate in real time will have limits on its knowledge and resources. It would be better to talk about "limits" on knowledge and resources, since "insufficient" can only be defined relative to some task, some environment, and some performance measure. What is insufficient for one task may be perfectly sufficient for another, and intelligence should be defined over a wide range of tasks.

3. An alternative definition of intelligence

Consider, then, an alternative approach based on the idea of an *agent*. Key elements of an agent are that it is situated in an environment, that it has limited knowledge, memory, computational capacity, and abilities to perceive and act in its environment, that it chooses its actions moment-to-moment, and that it has goals. An agent may have multiple goals simultaneously, and the goals, the environment, and the agent's capacities and abilities may evolve over time. We can call an agent *intelligent* if its moment-to-moment choices do, over time, lead it toward its goals, and if over time it can learn and adapt by increasing its knowledge and its ability to choose actions that lead it toward its goals.

Given this concept of an agent and what it means for an agent to be intelligent, we offer the following alternative definition of *intelligence*:

> *Intelligence is the ability of an agent, whether human, animal, artificial, or something else, to act in its environment in real time, using its limited knowledge, memory, computational power, and perception and action capabilities, choosing actions at each moment that move it toward its current goals, and to adapt over time by improving this ability to act.*

Central to this idea of intelligence is that an agent makes choices on a moment-to-moment basis, that the abilities and capacities to make these choices are limited, and that choices are made to move the agent in the direction of its goals. We would specifically exclude from being intelligent agents computer programs that always produce a certain predetermined output for a given input, systems whose only actions are to categorize the current input, even if the categorization was learned, or systems whose output is primarily determined by some random process that is independent of any perception of the environment.

4. Defining artificial intelligence

Given this definition of intelligence, we can move on to define *artificial intelligence*, in its two senses. In the AI1 sense, it is easy to say that artificial intelligence is the quality of intelligence in a man-made system. In its AI2 sense, AI is the field of study which considers how to design, construct, and evaluate AI1 systems. Now consider how to relate these definitions to some of the ideas in the field.

Consider the question of what constitutes an intelligent artificial agent. Silver et al. (2017) claim that the program they call AlphaGo Zero achieves "superhuman performance" starting "*tabula rasa*" with no human input. However, Marcus (2018) points out some problems with this claim. The deep neural network may learn without labeled input data, but this is only a small part of the whole system. Other parts of the complete agent were hand crafted by human experts, so the agent's performance as a whole actually depends on encoding a lot of human expertise. Thus if we consider

an entire agent that acts intelligently in the world, we have a different perspective on AI than when we focus on only a single component, however amazing its performance may be.

5. The Field of AI

From this perspective the question arises of what is the appropriate relationship between the field of AI, AI2, and the study of human intelligence. A prominent textbook (Russell and Norvig, 2010) begins on page 2 with a diagram showing that AI could involve thinking and acting "humanly" as well as thinking and acting "rationally." After further defining these terms, on page 5 they say that the rest of the book will focus just on the "acting rationally" quadrant, dismissing any consideration of modeling human intelligence. Their definition of rationality says that it is "an *ideal* performance measure," thus dismissing from the outset any consideration of human intelligence or the kinds of limitations we have included in our definitions.

Although not all researchers in the field will agree with this approach, it does exemplify the fact that much of AI research today ignores both human intelligence and Herb Simon's (Simon, 1996) concept of "bounded rationality," which takes into account limitations on knowledge and resources. Wang's emphasis on AIKR makes a very important point.

Laird, Lebiere, and Rosenbloom (2017) suggest a different approach. They discuss the concept of "*humanlike minds*," and even propose a "standard model of the mind" based on many years of research in cognitive architectures, in turn informed by research in psychology and cognitive science. Their approach to AI1 exemplifies a part of AI2 that does explicitly consider agency, human cognition, and limits or bounds on rationality. Such an approach fits much better with the definition of artificial intelligence we propose here. Since humans are the only instantiation of full general intelligence we know of today, it seems wise to consider an understanding of human intelligence as we search for better ways of creating artificial intelligence.

References

Laird, J. E., Lebiere, C., and Rosenbloom, P. S. 2017. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4):13.

Marcus, G. 2018. Deep Learning: A Critical Appraisal. `http://arxiv.org/abs/1801.00631`.

Russell, S. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, third edition.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–359.

Simon, H. A. 1996. *The Sciences of the Artificial*. Cambridge, MA: The MIT Press, third edition.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 50-51, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-07
Accepted 2019-11-07

Why Is Defining Artificial Intelligence Important?

Tomáš Mikolov TMIKOLOV@FB.COM
Facebook AI
San Francisco, California, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

The topic of defining artificial intelligence is a highly interesting problem for the research community, although it is not always obvious what can be gained from a new, or more precise definition. After all, many interesting definitions of AI have been provided in the past. As discussed in (Wang, 2019), the usefulness of a good definition can be in directing future research towards scientific problems that can be solved in the foreseeable future and that can lead to the development of a new and useful technology.

Interestingly, although AI research is very popular nowadays even in the mainstream media, a vast majority of the research efforts in academia and industry are directed towards applications of a known technology, with focus on minor and incremental improvements that do not have the ambition to lead to a major improvement of the AI approaches. In some sense, we are still living through the AI winter: although the popularity of the term 'AI' has increased greatly over the last years, a vast majority of the AI projects aim to solve very narrow, isolated tasks, with very limited efforts to define projects aiming on developing Human-level AI or AGI.

For these reasons, I find the above paper very interesting, as it provides a brief overview of the history of AI research from which it is clear that we are nowhere near the ultimate goal of developing machine intelligence, and that the ambitions of the AI community used to be much higher. On the other hand, I think some parts of the paper could be expanded. For example, to increase the fruitfulness of the provided definition of AI, it would be good to be more specific about which directions of AI research could change and how, and what problems the researchers should focus on more.

Even further, it would be good to discuss which important tasks might be solved if we redirect our attention in the proposed way. To give an example, much of the attention that the deep learning community has gained during the last decade comes from measurable advances of technology that is nowadays part of products used by billions of people: machine translation, image classification, speech recognition and so on. While it is expected that the development of novel approaches to AI can take many years, it could be good to have in mind tasks that may be one day solved if we guide our research using a new definition of AI.

Furthermore, although the problem of defining AI is a very difficult one, it might be considerably easier to define what AI is not. The paper currently discusses this very briefly, giving examples such as programs that perform far too obvious function (sorting of numbers, or a basic calculator). I think this part of the paper could be expanded, as it would be later easier to argue that there is indeed a need for a good definition of what AI is to stimulate novel research.

I also think that a definition of 'Useful AI' might be relevant here. After all, the paper concludes that a good definition of AI is "adaptation with insufficient knowledge and resources," which sounds more like a definition of artificial life as it does not consider human users of the AI. If we assume that from the set of all possible AIs, we are the most interested in those that actually produce something useful for its human users, we should have in mind the utility of the system. We may then define the 'Useful AI' as such a computer system that requires the least amount of human intervention and physical time to adapt to perform a new, useful task.

References

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 52-56, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-11
Accepted 2019-11-07

What Is Artificial Intelligence?

William J. Rapaport RAPAPORT@BUFFALO.EDU
Departments of Computer Science and Engineering, Philosophy, & Linguistics,
and Center for Cognitive Science
University at Buffalo
The State University of New York
Buffalo, New York, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Wang (2019) claims to define AI in the sense of delimiting its research area. But he offers a definition only of 'intelligence' (not of AI). And it is only a *theory* of what intelligence is (artificial or otherwise). I offer and defend a definition of AI as computational cognition.

1. The Nature of Definitions

Forward-looking (or *prescriptive*) definitions suggest what researchers ought or want to work on. Backward-looking (or *descriptive*) definitions are based on what researchers have actually tried to do or succeeded in doing.[1] Examples abound in the history of science: Atoms were originally conceived as indivisible; we now know that they are not; electrons were originally conceived as tiny particles, a view now challenged by quantum mechanics. Reconciling such apparently incommensurable definitions or concepts is an open question in philosophy.

In the case of AI, there is an obvious candidate for the forward-looking, prescriptive definition (mentioned, but not explicitly cited, by Wang): McCarthy's definition from the Dartmouth conference that gave the field its name:

> the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. (McCarthy et al., 1955)

This definition is free from the anthropomorphism that Wang criticizes in others (although McCarthy et al. do go on to talk of solving "problems now reserved for humans").

Whether modeled on humans or not, AI has also tended to be Janus-faced, with the interaction between the naturally-occurring original and its computational model going in both directions, as in these two definitions:

1. . . . *artificial intelligence*, the science of making machines do things that would require intelligence if done by men. (Minsky, 1968, p. v)
2. By "artificial intelligence" I . . . mean the use of computer programs and programming techniques to cast light on the principles of intelligence in general and human thought in particular. (Boden, 1977, p. 5)

1. "The dictionary, after all, is more of a rearview mirror than a vanguard of change"—Peter Sokolowski, cited in Fortin (September 20 2019).

Here, the anthropomorphism is surely eliminable (delete "if done by men" from Minsky's, and "and human thought in particular" from Boden's). Minsky looks at naturally occurring "intelligence" and seeks to re-implement it in machines. Boden looks at computation and seeks to use it to understand "intelligence".

And, of course, there are problems (noted by Wang) raised by the "fluidity" of concepts and the difficulty (if not impossibility) of providing necessary and sufficient conditions for concepts best understood as having only family resemblances. As a consequence, one-sentence definitions such as any of those under discussion are really only acceptable for quick overviews or dictionaries. To really understand a subject, one needs at least an encyclopedia article, a textbook, or a research program (Rapaport, 2019, §3.3.3).

2. Wang's Definition

"On Defining Artificial Intelligence" offers no such definition. Ignoring 'A', Wang concentrates on 'I': "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources" (p. 17). As definitions of intelligence go, this is not a bad one, though I find it notable that Wang pays scant attention to definitions of intelligence from the psychological literature (e.g., Gardner 1983; Sternberg 1985). Moreover, Bertram Raphael observed "that AI is a collective name for problems which we do not yet know how to solve properly by computer" (Michie, 1971, p. 101), which implies that, once we do know how to solve them, they are no longer AI (Wang, 2019, p. 11). Daniel R. Schlegel (personal communication) points out "Without the 'capacity' part of ... [Wang's] definition, this would be lurking in his definition—once something is understood to the point that adaptation is no longer required, it isn't an intelligent action anymore."

What about 'A'? Wang says that he won't discuss the possible confusion with 'artificial' in the sense of "fake" (p. 3) and that "how to interpret the 'A' is not a big issue" (p. 4). I think this is a mistake. The nature of AI's "artificiality" has played an important role in philosophical discussion: The argument from biology in Searle (1980) states in essence that an AI that is A is therefore not I (Rapaport 2000b; Rapaport 2019, §19.6.2).

Wang suggests that his definition of intelligence "corresponds to a *working condition* and a *coping strategy* that are both different from those of computation" (p. 17). If so, then what does AI's artificiality consist in? Yet he suggests that AI both will and will not be algorithmic:

> ... an intelligent system defined in this way cannot always solve problems by following problem-specific algorithms .... On the other hand, a computer system eventually runs according to algorithms. The solution of this dilemma is to combine algorithm-specified steps to handle each problem-instance in a *case-by-case* manner .... (p. 20)

He seems to think that if AI is computational, then there must be a *single* algorithm that does it all (or that is "intelligent"). He agrees that this is not possible; but whoever said that it was?

He also puts a lot of weight on the view that "A program is traditionally designed to do something in a predetermined *correct* way ..." But AI researchers from the very beginning have relied on "heuristics", not in the sense of vague "rules of thumb" or fallible suggestions of how to do something, but in a very precise *algorithmic* sense:

> A *heuristic for problem p* can be defined as an *algorithm* for some problem p', where the solution to p' is "good enough" as a solution to p (Rapaport, 1998, p. 406). Being "good

> enough" is, of course, a subjective notion; Oommen and Rueda (2005, p. 1) call the "good enough" solution "a *sub-optimal* solution that, hopefully, is arbitrarily close to the *optimal*." (Rapaport 2017, p. 15; Rapaport 2019, §3.15.2.3; see also Romanycia and Pelletier 1985; Chow 2015)

Thus understood, an AI heuristic is a "predetermined correct way" to do something that is (arbitrarily) *close to* what minds do. It is related to Simon's notion of bounded rationality; so (given Wang's remarks in §4.1), Wang should be sympathetic to it.

As for his comment that

> traditional computer systems should be taken as unintelligent, as they are designed according to principles that are fundamentally different from what we call intelligence. From a theoretical point of view, AI should not be considered as the same as computer science, or a part of it. (p. 16)

one should consider the fact that Turing Machines themselves were conceived along the lines of McCarthy's and Minsky's methodology: Analyze how humans solve a certain problem, and then devise an algorithm that does the same thing in the same way (Rapaport, 2017, p. 12).

3. My Definition

AI is a branch of computer science (CS), which is the scientific study of what problems can be solved, what tasks can be accomplished, and what features of the world can be understood computationally (i.e., using the language of Turing Machines), and then to provide algorithms to show how this can be done efficiently, practically, physically, and ethically (Rapaport 2017, p. 16; Rapaport 2019, §3.15). Given that CS's primary question is "What is computable?", I take the focus of AI to be on whether *cognition* is computable.

I agree with Wang that both 'A' and 'I' are not the best terms, so I replace 'A' by 'computational' and 'I' by 'cognition': Computational cognition (which we can continue to abbreviate as 'AI') is the branch of CS that tries to understand the nature of cognition (human or otherwise) computationally. By 'cognition', I include such mental states and processes as belief, consciousness, emotion, language, learning, memory, perception, planning, problem solving, reasoning, representation (including categories, concepts, and mental imagery), sensation, thought, etc. AI's primary question is "How much of cognition is computable?"; its working assumption is that *all* of cognition is computable (echoing McCarthy's original definition); and its main open research question is "Are *aspects of cognition that are not yet known to be computable* computable?" If they are, does that mean that computers can "think" (i.e., produce cognitive behavior)? If there are *non*-computable aspects of cognition, *why* are they non-computable? An answer to this question should take the form of a logical argument such as the one that shows that the Halting Problem is non-computable. It should not be of the form: "All computational methods tried so far have failed to produce this aspect of cognition". After all, there might be a new kind of method that has not yet been tried.

Wang's definition of intelligence is a proposal about *how to go about finding* computational solutions to cognitive abilities. Do any of those solutions also need to be solutions to the problem of how *living* entities cognize? *Pace* Boden, not necessarily, for at least two reasons. First, a process is comput*able* iff there is an algorithm (or perhaps multiple interacting algorithms) that is input-output equivalent to the process. There is no requirement that natural entities that exhibit a

comput*able* behavior must themselves do it computa*tionally* (Rapaport, 1998, 2012, 2018). Second, as Shapiro (1992)[2] has urged, there are 3 distinct goals of AI: (1) AI as advanced computer science or engineering extends the frontiers of what we know how to program and to do this *by whatever means will do the job*, not necessarily as humans do it. (2) AI as computational psychology writes programs as theories or models of *human cognitive behavior*. (3) AI as computational philosophy investigates *whether cognition in general* (and not restricted to *human* cognitive behavior) *is computable*.

Wang has two objections to defining AI as computational cognition. First, he suggests that some of the items included under cognition as characterized here are simply "other vague concepts" (p. 5), themselves in need of definition. But my proposal first refines 'I' to 'cognition', and then further refines 'cognition' to that (family resemblance) list above. Refining those further becomes one of the tasks of AI (along with the other cognitive sciences). To the extent that AI succeeds, each aspect of cognition will be made precise.

Second, Wang raises the specter of "fragmentation" (p. 12): separate solutions to each aspect of cognition, but no unified one such as we humans apparently have. This problem does need to be addressed: Various modes of cognition do have to interact somehow, but it doesn't follow that a single AI "master algorithm" is needed. Separate modules with a central coordinating system is also a possibility. Fragmentation in other sciences, such as math or physics, has not been a serious obstacle to progress.

References

Boden, M. A. 1977. *Artificial Intelligence and Natural Man*. New York: Basic Books.

Chow, S. J. 2015. Many meanings of 'heuristic'. *British J. Phil. Sci.* 66:977–1016.

Fortin, J. September 20, 2019. When dictionaries wade into the gender (non)binary. New York Times. https://www.nytimes.com/2019/09/20/style/they-nonbinary-dictionary-merriam-webster.html.

Gardner, H. 1983. *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. 1955. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth.html.

Michie, D. 1971. Formation and execution of plans by machine. In Findler, N. and Meltzer, B., eds., *Artificial Intelligence and Heuristic Programming*. New York: American Elsevier. 101–124.

Minsky, M. 1968. Preface. In Minsky, M., ed., *Semantic Information Processing*. Cambridge, MA: MIT Press. v.

Oommen, B. J. and Rueda, L. G. 2005. A formal analysis of why heuristic functions work. *Artif. Intell.* 164(1-2):1–22.

Rapaport, W. J. 1998. How minds can be computational systems. *J. Exp. Theor. Artif. Intell.* 10:403–419.

2. See also (Rapaport, 1998, 2000a, 2003).

Rapaport, W. J. 2000a. Cognitive science. In Ralston, A., Reilly, E. D., and Hemmendinger, D., eds., *Encyclopedia of Computer Science*. Fourth edition. 227–233.

Rapaport, W. J. 2000b. How to pass a Turing test: Syntactic semantics, natural-language understanding, and first-person cognition. *J. Logic, Lang., & Info.* 9(4):467–490.

Rapaport, W. J. 2003. What did you mean by that? Misunderstanding, negotiation, and syntactic semantics. *Minds and Machines* 13(3):397–427.

Rapaport, W. J. 2012. Semiotic systems, computers, and the mind: How cognition could be computing. *Int'l. J. Signs & Semiotic Systems* 2(1):32–71.

Rapaport, W. J. 2017. What is computer science? *Amer. Phil. Assn. Newsletter on Phil. & Computers* 16(2):2–22.

Rapaport, W. J. 2018. Syntactic semantics and the proper treatment of computationalism. In Danesi, M., ed., *Empirical Research on Semiotics and Visual Rhetoric*. Hershey, PA: IGI Global. 128–176.

Rapaport, W. J. 2019. Philosophy of Computer Science. `http://www.cse.buffalo.edu/~rapaport/Papers/phics.pdf`.

Romanycia, M. H. and Pelletier, F. J. 1985. What is a heuristic? *Comp. Intel.* 1(2):47–58.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417–457.

Shapiro, S. C. 1992. Artificial Intelligence. In Shapiro, S. C., ed., *Encyclopedia of Artificial Intelligence*. New York: John Wiley & Sons. 54–57.

Sternberg, R. J. 1985. *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge, UK: Cambridge University Press.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 57-59, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-17
Accepted 2019-11-07

On Pei Wang's Definition of Artificial Intelligence

Raúl Rojas ROJAS@INF.FU-BERLIN.DE
Dept. of Mathematics and Computer Science
Freie Universität Berlin
Berlin, Germany

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

There has been a long discussion in the research community, spanning several decades about the definition of the term "Artificial Intelligence." Pei Wang's paper (Wang, 2019) reviews the most relevant contributions to this debate, comparing their strengths and weaknesses. He then proposes the following definition of AI: "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources."

1) The machine dimension. I am a little surprised that the definition attempts to cover biological and artificial systems simultaneously. I would have expected a definition of "intelligence" that applies to living beings first. If we then say "artificial intelligence," it is clear that we mean "what biological systems can do, now done by computers." The really big issue is to outline intelligence in biological systems and then discover its essential ingredients. Connecting the definition to computers (the artificial) is then rather straightforward (see my fifth comment below).

The limits of what an "information-processing system" could be are only implicit in Wang's definition. The earth, for example, can be regarded as one such system if we categorize physical interactions as information. But the earth does not have any knowledge, not even insufficient knowledge, thus it is clear that the definition can only apply to *individual* living beings.

2) The historical dimension. I think that the paper shows how the definition of AI has been changing across authors. However, it has also been changing drastically across time. What we can observe is that computers have been able to perform activities, or have acquired abilities originally characteristic of humans. Just doing simple arithmetic is one example. Until the arrival of the computer only persons could multiply or divide. When the first computers were built, some of them were called "electronic brains," like the American ENIAC in 1945. Even today, it seems that in China people refer colloquially to the computer as an electronic brain (電腦).

This means that AI has a historical component that we cannot disregard. Computer algebra is one example. When I started working in the field of AI in the 1970s, one of my first projects was writing a computer algebra system. Computing derivatives, integrals and solving equations symbolically, was something computers were just starting to do. AI programming languages, such as Lisp, were used to write symbolic manipulation code and pattern matchers in order to solve algebraic problems. Today, computers are much better at manipulating conventional algebraic expressions than humans are. They are in fact superhuman-algebraists. Thus, research in this field has moved from the AI quarters to the offices of scientific computing experts.

Another example is chess. Almost all books about AI used to start discussing search algorithms, and chess was a good example of what you can do if the computer is able to rapidly inspect deep

decision trees. Chess was the paradigmatic AI project for many years, and it is not surprising that Alan Turing, Claude Shannon and the German inventor Konrad Zuse, all of them early pioneers of computing, were very interested in automating the game. Nowadays, very few people in the academic community are interested in chess. The only exception is when a new approach is tested in order to compare its results with the reigning chess programs (for example, by applying reinforcement learning so that the computer learns from playing against itself).

What I want to stress then is that AI research is a moving frontier: many things we called AI yesterday are not considered AI today, because we know how to solve those problems.

3) The biological dimension. Wang's definition makes clear that we perceive intelligence whenever there is adaptation. If an insect feels pain and moves away from a chemical, we consider that to be intelligent behavior, one developed by the animal during an evolutionary process. Animals that could not adapt have just disappeared from the face of the earth.

However, we can detect adaptive behavior even in plants. It is now well known that plants have sensors for detecting different wavelengths of photons, and chemicals diffused through the atmosphere. In forests, trees can exchange chemical signals through their roots, so that, surprisingly, a forest is actually a gigantic information processing ensemble, where the individual plants are generating and consuming information. Can we then talk about "plant intelligence"? There are researchers who refer to animal as well as to plant intelligence.

In my opinion, intelligence is a continuum. If we grade humans with intelligence of 1.0, maybe we can grade other primates with 0.9, crows with 0.85, and so on. A small bacterium can adapt to its environment, build colonies, move towards food or light. Even a single cell already shows not just adaptation but what we would call "purposeful behavior" to some extent. Animal intelligence is a reality and the only problem is where to *draw the line*: is bacterial intelligence a 0.0, or is it greater than zero?

4) The phase transition dimension. The problems mentioned in commentary (3) all arise because we have to make a distinction between information and knowledge. Information can be just physics. Astronomers study black holes and how they swallow information. Information is exchanged on the forest floor. But at what level do we start to talk about knowledge? The word means implicitly that someone "knows‘," but the question is whether those someones are aware that they know. Obviously, a plant is not aware of anything, although it possesses some kind of precursor to a nervous system. But is an insect aware of a noxious chemical so that we could say the insect has to apply "knowledge"? If we set the knowledge bar too high, then few animals would be called intelligent. The rest would be just automata, as René Descartes suggested centuries ago.

I think that in the continuum of adaptive behavior from 0.0 to 1.0 there are several "phase transitions" that are important for intelligence at the human level. Let me mention four: the ability to feel pain, to be aware, to feel emotion, and to reason.

I have often carried out the following experiment with students at my university. I provide them with a list of animals, from simple bacteria and insects, reptiles, mammals, primates, all the way to humans. I then ask them to draw a line where they think that the animals do not have one of the four capabilities listed above. For "pain" they usually draw the line between plants and insects. Plants do not feel pain, insects do. For awareness they draw the line between insects and fish/reptiles. Insects do not seem to be aware of anything, while we trust fish and reptiles with some kind of "knowledge" about their place in the world. For emotions, most students draw the line after reptiles, meaning that birds and mammals can feel them. A sad or happy reptile seems to be something that we cannot imagine. The last phase transition is achieving reasoning, a unique human capability.

Human intelligence seems to require all four phase transitions. Antonio Damasio (1994) has made a powerful case about the importance of our body and our emotions for displaying the full range of human intelligence. We are of course aware of our place in the world and we can feel physical and mental pain. Through reasoning we are also aware of our mortality, maybe the only species to do so. It can be argued that all religions and even philosophy are attempts to come to terms with this fact.

So, going back to Wang's definition: where do we draw the line between information and knowledge? If knowledge is awareness, then we draw the line earlier between species. If knowledge is a neighbor of reasoning, then we are really close to primates and humans.

5) The industrial dimension. It is unfortunately so that a definition of AI is not going to change the practice of doing AI. Normally what happens is that there is one application, for example, speech recognition, which is a human capability that we want to transfer to computers. Likewise recognition of images. Or driving cars. If only humans can do it up to now, and we want to do it with computers, then it is AI.

The largest research centers for AI are no longer at universities. They are run by Google, Facebook, Amazon, Apple, Microsoft, and IBM. AI applications can now be embedded in microchips so that a small gadget can recognize my spoken requests. A computer at the airport compares my face with the biometric data in my passport and lets me walk through.

AI systems are being used to make communication with the computer easier (using speech recognition), to model my consumption patterns, predict my "next move" in the browser, to offer me merchandise, retrieve information I might consider interesting, and so on. Today's AI is like a hydra: it is many applications running concurrently, doing different unconnected things in my computer, most of them applying either brute force search, or exact numerical algorithms.

I don't think that human and machine intelligence will converge. On the contrary, they are currently diverging. Human intelligence is based on pattern recognition, intuition and filtering of unnecessary details. Computer intelligence is based on fast electronics and optimal algorithms. In fact, many people dream of applying quantum computers to machine intelligence so that we could search in enormous spaces of possibilities in milliseconds. Computer intelligence will not require all the phase transitions that we humans need. Computers will never feel and will never have religion or philosophy because they cannot die.

Therefore, we really need two separate definitions. One for what biological intelligence is, and one for what artificial intelligence could be. With that said, we have to thank Pei Wang for starting a much-needed discussion with his paper.

References

Damasio, A. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. London: Picador.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 60-62, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-27
Accepted 2019-11-07

On Defining Artificial Intelligence—Commentary

Marek Rosa MAREK.ROSA@KEENSWH.COM
GoodAI
Prague, Czech Republic

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

1. Introduction

As Wang (2019) suggests, defining artificial intelligence (or more importantly, intelligence) is a crucial problem for researchers to grapple with to help them guide their work. It was a problem we addressed at GoodAI in order to develop our working definition of "intelligence," which has helped us layout our framework for AI (Rosa and Feyereisl, 2016).

At GoodAI we defined intelligence as a "problem-solving tool that searches for solutions to problems in dynamic, complex and uncertain environments" (Rosa and Feyereisl, 2016). This can be simplified further by viewing most problems as search and optimization problems (Polya, 1971), where the goal of intelligence is to always find the best available solutions with as few resources and as quickly as possible (Gershman, Horvitz, and Tenenbaum, 2015; Marblestone, Wayne, and Kording, 2016).

I believe that our definition fits well with Wang's definition of intelligence as "adaptation with insufficient knowledge and resources," and also embodies the three Assumptions of Insufficient Knowledge and Resources (AIKR) outlined by Wang. Below I build on some of the ideas in the paper and compare them to the work we are undertaking at GoodAI.

2. Refining the definition

When defining AI, we started with a broad definition of intelligence (as stated above) and this helped us to direct our research methods. When we view intelligence as a search process that helps us narrow down the search space, invent new search skills, and generate relevant hypotheses, and the development of AGI is also viewed as search itself, we turn towards methods that help us leverage various search principles that embody such biases and prior knowledge about the search process. These methods include: meta-learning, multi-agent learning, adversarial learning, evolution-inspired search, or search without objectives (Lehman and Stanley, 2011). We do this as we believe that these might be suitable tools for automating the search and narrowing down the large search space of possibilities as effectively as possible.

When we want to start measuring the success of our agents we need to be more specific with our definition. As Wang mentions, intelligence has many grades and shades and not all evaluation tests are applicable to all agents (Hernández-Orallo, 2017). For example, we might have a good "baby-AGI" algorithm that fails on tasks if they're not given to it in the right order of complexity.

We believe there are two components to developing AGI: a core meta-algorithm, together with a structured and guided learning process (e.g. a curriculum, self-play, etc.). So, we look at the instrumental definition from the point of view of tasks and skills: we define specific skills that we want to test on concrete tasks in a learning curriculum. Being able to complete these skills helps us define whether an agent is intelligent.

3. Minimal learning environment

Our instrumental definition has guided our aim of creating a minimal necessary learning environment in which AI agents can demonstrate intelligence. We are still working to define the minimal set of tasks that show intelligence, but we consider graduality, or the re-use of previously learned knowledge to solve new tasks, and meta-learning, as cornerstones. Once an agent solves the set of tasks, and at the same time passes tests proving that the agent was reusing previously learned skills, we would consider it to be intelligent (or at least reaching a certain degree of minimal intelligence).

Humans are equipped with the same intelligence to solve modern-day tasks, however biologically we're no different from our ancestors thousands of years ago (Kralik, 2018). Theories of why behavioral modernity emerged could help us create the right minimal necessary learning environment and tasks.

We believe that an instrumental definition of intelligence can be built up from the necessary skills and abilities that an intelligent being should possess. In our Roadmap (Rosa and Feyereisl, 2016) we identify intrinsic learning skills, including three intrinsic core skills, namely gradual learning, guided learning, and learning to learn. We developed tests to check whether an agent uses these meta-skills while progressing in a curriculum (described in more detail below), such as tests for graduality and avoiding catastrophic forgetting.

4. Curriculum learning

Our definition led us to create a list of intrinsic learning skills and associated tasks structured as a learning curriculum, which we believe if an agent could learn from and solve, it would be displaying higher-levels of intelligence (Rosa and Feyereisl, 2016). This process helped us realize and distill not only the core principles and skills that an AI system should possess but also the systems' subsequent evolution. This is still an ongoing process, one that would be impossible without a working definition of AI.

Like Wang, we are not as concerned with whether agents can complete a task in the predetermined "correct" way, but the key metric is how fast it adapts and learns to solve novel tasks (using its general problem-solving and learning skills). In addition, we designed specific tasks in the curriculum in which the agent has to reuse previously learned skills on unseen tasks.

We agree with Wang that environments and agents change with time, so solutions to problems cannot always be replicable. Therefore, a system cannot rely on algorithms for specific problems. "Instead, it should focus on the design of the algorithmic steps as the building blocks of problem-solving processes, as well as on the mechanism to combine these steps at run time for each individual problem-instance" (Wang, 2019, p. 20). Correspondingly, we also see adaptation as the key metric for success and hence a key ingredient to the definition of intelligence.

Although Wang argues that definitions of intelligence should not be over anthropocentric, we believe that teaching AI to understand our world and how to communicate using human-

interpretable-language is also a vital task. Although communication may not be an intrinsic part of intelligence, it could be an extremely useful tool in building and understanding intelligent systems. Communication with our agents could help avoid ambiguities, and having agents with a deep understanding of our world could have a direct impact on AI safety.

5. Conclusion

This paper comes at an important time when the phrase artificial intelligence is being used all around us for very different things and purposes. For example, most narrow AI systems which are trained by an engineer or researcher will not adapt once deployed, this is very different from a general AI system which we believe must have the ability to adapt. One of the key goals of our Framework in 2016 was to create a "unified collection of principles, ideas, definitions, and formalizations of our thoughts on the process of developing general artificial intelligence." The idea being that this would help researchers communicate better and push towards a common goal. As Wang identified, by splintering the field into various perspectives researchers begin their journeys up different mountains rather than pushing together to reach the same summit.

We agree with Wang's assessment that the three Assumptions of Insufficient Knowledge and Resources are normal working conditions of intelligent systems and along with the ability to adapt, should be present in any definition of artificial intelligence. Therefore, we also agree that there is no one "true" definition of AI but that some definitions are more useful than others.

References

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349(6245):273–278.

Hernández-Orallo, J. 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press.

Kralik, J. D. 2018. Core High-Level Cognitive Abilities Derived from Hunter-Gatherer Shelter Building. In Juvina, I., Houpt, J., and Myers, C., eds., *Proceedings of the 16th International Conference on Cognitive Modeling*, 49–54. Madison, WI: University of Wisconsin.

Lehman, J. and Stanley, K. O. 2011. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19(2):189–223.

Marblestone, A. H., Wayne, G., and Kording, K. P. 2016. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience* 10:94.

Polya, G. 1971. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press.

Rosa, M. and Feyereisl, J. 2016. A Framework for Searching for General Artificial Intelligence. arXiv:1611.00685 [cs.AI]. `http://arxiv.org/abs/1611.00685`.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 63-65, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-30
Accepted 2019-11-07

A Broader, More Inclusive Definition of AI

Peter Stone PSTONE@CS.UTEXAS.EDU
Department of Computer Science
University of Texas at Austin
Texas, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Pei Wang's article "On Defining Artificial Intelligence" (Wang, 2019) is a thoughtful and well-written argument in favor of a particular working definition of AI and an associated research project called NARS, for Non-Axiomatic Reasoning System.

It is structured in three parts. First, it argues in favor of the need to define Artificial Intelligence. Second, it argues in favor of Wang's particular definition. Third, it explains how his definition leads to NARS.

I focus my comments here on the first and second parts. While there is much to agree with in the article, in the interest of discourse, I further focus my comments on the points with which I disagree. That being said, I fully acknowledge that it is much easier to criticize than it is to write an article with no room for criticism!

1. A Much Broader Definition

From the highest level perspective, I agree with Wang's exposition of the values of specifying one's working definition, and commend him for acknowledging on more than one occasion that there is room for different definitions. But despite this acknowledgment, I note that on more than one occasion he seems to argue for the need to converge on a single definition or the superiority of his own definition, neither of which I endorse. Personally, I hold strongly to the "big tent" view of AI that allows, and even encourages, multiple perspectives and agendas, and thus working definitions, to co-exist within the same field. It is with this view that I prefer a broad definition such as the one we put forth in the 2016 report of the One Hundred Year Study on AI (Stone et al., 2016):

> "Artificial Intelligence (AI) is a science and a set of computational technologies that are by inspired by—but typically operate quite from differently from—the ways people use their nervous systems and bodies to sense, learn, reason, and take action."

2. The need for a Definition

Section 1.3 of the article presents a very useful discussion on what a definition is, and Section 2 lays out an interesting classification and generalization of the various types of AI definitions.

However in my opinion, the article over-reaches in a few ways. For example:

> "Though a well-defined concept is not easy to obtain, its benefits are hard to overstress. It will prevent implicit assumptions from misguiding a research project."

While having a definition may indeed help focus or guide one's research, I would not go so far as to say that research is "misguided" if it is not tied closely to a particular definition of AI.

Even for research that does start from a definition, Wang writes:

> "In particular, the definition distinguishes the features of human intelligence that need to be reproduced in an AI system from those that can be omitted as irrelevant."

In my opinion, this statement leads directly to the need for a plethora of working definitions, so that as a field we can investigate a broad range of the features of artificial intelligence. In fact, the statement requires different working definitions over time as knowledge and tools progress. Just as there are now considered to be different types of human intelligence (e.g. spatial intelligence, emotional intelligence, etc.), the field of AI has room for, and indeed requires, investigations of machine intelligence from various perspectives.

A few more minor points that bear mentioning follow.

- One justification raised for needing a single definition is so that policy makers can assess what AI systems will be able to do in the future. On the contrary, I think it is incumbent on AI researchers to stress that AI is *not one thing* and should therefore not be regulated as such. Policies ought to be developed sector by sector with regards to specific AI-based technologies that are relevant to that sector (see the AI100 report for further discussion on this point).
- Another justification put forth for a definition of AI is so that we will know "how to build one." I disagree that AI is a "thing" to be built, and again it is certainly not *one* thing.
- The definition of Capability-AI takes an applications-oriented perspective, but then seemingly limits AI research in this paradigm to *matching* human performance. It ought to leave room for superhuman performance being realized by AI-based systems, as we have seen from recent game-playing systems.

In summary, I agree with the author regarding the usefulness of working definitions for helping focus one's research. But I caution that definitions can also be exclusionary, and object to attempts to use narrow definitions as justification for limiting the field by dictating what "counts" as AI. The author is correct that the inclination to coin terms such as "AGI" has arisen to differentiate from AI research that is more narrowly focused. However, I disagree with the need to differentiate in this way. The term AI, and the field of AI can, and do, encompass both narrow and broad research foci and applications.

3. Wang's Definition

As for Wang's working definition itself, I think it is perfectly fine as "a" definition of AI. However I do not endorse it as "the" definition.

Actually, I do not find that the definition stands alone. Rather, to fully understand it requires reading its explanation throughout the 2.5 pages of Section of 3.2. For example, the phrase "adapt to its environment" does not necessarily lead to any requirements over beliefs, actions, tasks, or problems. And the meaning of "insufficient knowledge and resources" only becomes clear through the prose that follows. This need for extensive explanation violates at least the "exactness" desideratum of a good definition.

In any case, the author's definition is very well-paired with his research program and vice versa. It is an elegant coupling that is indeed commendable and worthy of emulation. However the leap from there to statements such as the following goes too far.

> "The current field of AI is actually a mixture of multiple research fields, each with its own goal, methods, applicable situations, etc., and they are all called AI mainly for historical, rather than theoretical, reasons."

In my opinion, the field of AI can tolerate, and in fact actively benefits from, research projects and perspectives that arise from a variety of working definitions, or that are even not directly tied to working definitions at all. It is for that reason that for the purpose of defining the field, I much prefer the AI100's much broader definition, as quoted above.

References

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., and Teller, A. 2016. Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence. Stanford University, Stanford, CA. `http://ai100.stanford.edu/2016-report`.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 66-67, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-27
Accepted 2019-11-04

John McCarthy's Definition of Intelligence

Richard S. Sutton RSUTTON@UALBERTA.CA
University of Alberta
Edmonton, Alberta, Canada

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Pei Wang (2019), in the target paper, is right to stress the importance of a scientific field having a generally agreed on definition of its subject matter. He is also right when he says that many artificial intelligence (AI) researchers accept, in their textbooks and public statements, that there is no satisfactory way to define intelligence. However, for other AI researchers—including me—this is not acceptable. A field needs to be able to reason, at least in a general way, from a clear statement of its subject matter.

But is there really no standard definition of intelligence within AI? Actually, it is not hard to find a public statement by a prominent AI researcher defining intelligence. The definition given by John McCarthy (1997), the AI researcher who coined the phrase "artificial intelligence," is:

> "Intelligence is the computational part of the ability to achieve goals in the world."

I find this simple and commonsense definition to be useful and satisfying, although it is not specifically mentioned in the target paper.

According to McCarthy's definition, intelligence is an ability, and so of course a system may possess that ability to various degrees. Thus the definition does not make an absolute distinction between systems that are intelligent and those that are not. A person, a thermostat, a chess-playing program, and a corporation all achieve goals to various degrees and thus can be thought of as intelligent to those degrees. This is just as it should be, in my opinion.

McCarthy's definition also specifies that intelligence is the computational part of that ability, ruling out, for example, systems that achieve their goals merely by being physically strong, or by having superior sense organs.

At the heart of McCarthy's definition is the notion of "achieving goals." This notion is clear, but informal. What does it mean, exactly, to have a goal? How can I tell if a system really has a goal rather than just appears to? These questions seem deep and confusing until you realize that a system having a goal or not, despite the language, is not really a property of the system at all. It is a property of *the relationship between the system and an observer*. It is a 'stance' that the observer takes with respect to the system (Dennett, 1989). The relationship between the system and an observer that makes it a *goal-seeking* system is that the system is most usefully understood (i.e., predicted or controlled) by the observer in terms of the system's *outcomes* rather than in terms of its *mechanisms*.

For example, for a home owner, a thermostat is most usefully understood in terms of its keeping the temperature constant—an outcome—and thus for the home owner *the thermostat has a goal*. But for a repairman fixing a thermostat, it is more useful to understand the thermostat at a more

mechanistic level—and thus for the repairman *the thermostat does not have a goal*. The thermostat either does or does not have a goal depending on the observer, depending on whether the outcome view or the mechanism view of the thermostat is more useful. Even for a single observer, which view is more useful may change over time, and thus the same system may change from not having a goal to having one (or vice versa), as when the thermostat repairman fixes his own home's thermostat using the mechanism view, and then uses the thermostat to control the temperature of his house using the outcome view. And of course there may be degrees to which the two views are useful, and thus degrees of goal-seeking-ness. As in the case of intelligence itself, the notion of having a goal or not is not an absolute dichotomy, but a question of degree.

Another good example of goal-seeking-ness varying with the observer is that of a computer chess program. Suppose I am playing the program repeatedly. If I don't know how it works and it plays better than I, then my best understanding of the program is probably that it has the goal of beating me, of checkmating my king. That would be a good way of predicting the near-inevitable outcome of the games, despite how I might struggle. But if I wrote the chess program (and it does not look too deep), then I have an alternative mechanistic way of understanding it that may be more useful for predicting it (and for beating it).

Putting the two ideas together, we can define intelligence concisely and precisely:

> "Intelligence is the computational part of the ability to achieve goals. A goal achieving system is one that is more usefully understood in terms of outcomes than in terms of mechanisms."

References

Dennett, D. C. 1989. *The Intentional Stance*. MIT press.

McCarthy, J. 1997. What is Artificial Intelligence? Available electronically at `http://www-formal.stanford.edu/jmc/whatisai/whatisai.html`.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Journal of Artificial General Intelligence 11(2) 68-70, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-17
Accepted 2019-11-07

On Defining Differences between Intelligence and Artificial Intelligence

Roman V. Yampolskiy ROMAN.YAMPOLSKIY@LOUISVILLE.EDU
Computer Engineering and Computer Science
University of Louisville
Louisville, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

In "On Defining Artificial Intelligence" Pei Wang (2019) presents the following definition: "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources." Wang's definition is perfectly adequate and he also reviews definitions of intelligence suggested by others, which have by now become standard in the field (Legg and Hutter, 2007). However, there is a fundamental difference between defining intelligence in general or human intelligence in particular and defining Artificial Intelligence (AI) as the title of Wang's paper claims he does. In this commentary I would like to bring attention to the fundamental differences between designed and natural intelligences (Yampolskiy, 2016).

AI is typically designed for the explicit purpose of providing some benefit to its designers and users and it is important to include that distinction in the definition of AI. Wang only once, briefly, mentions the concept of AI safety (Yampolskiy, 2013; Yampolskiy and Fox, 2012; Bostrom, 2014; Yudkowsky, 2011; Yampolskiy, 2015a) in his article and doesn't bring it or other related concepts into play. In my opinion, definition of AI which doesn't explicitly mention safety or at least its necessary subcomponents: controllability, explainability (Yampolskiy, 2019b), comprehensibility, predictability (Yampolskiy, 2019c) and corrigibility(Soares et al., 2015) is dangerously incomplete.

Development of Artificial General Intelligence (AGI) is predicted to cause a shift in the trajectory of human civilization (Baum et al., 2019). In order to reap the benefits and avoid pitfalls of such powerful technology it is important to be able to control it. Full control of intelligent system (Yampolskiy, 2015b) implies capability to limit its performance (Trazzi and Yampolskiy, 2018), for example setting it to a particular level of IQ equivalence. Additional controls may make it possible to turn the system off (Hadfield-Menell et al., 2017), and turn on/off consciousness (Elamrani and Yampolskiy, 2019; Yampolskiy, 2018a), free will, autonomous goal selection and specify moral code (Majot and Yampolskiy, 2014) the system will apply in its decisions. It should also be possible to modify the system after it is deployed to correct any problems (Yampolskiy, 2019a; Scott and Yampolskiy, 2019) discovered during use. An AI system should be able, to the extent theoretically possible, explain its decisions in a human comprehensible language. Its designers and end users should be able to predict its general behavior. If needed, the system should be confinable to a restricted environment (Yampolskiy, 2012; Armstrong, Sandberg, and Bostrom, 2012; Babcock, Kramár, and Yampolskiy, 2016), or operate with reduced computational resources. AI should be operating with minimum bias, and maximum transparency, it has to be friendly (Muehlhauser and Bostrom, 2014), safe and secure (Yampolskiy, 2018b).

Consequently, we propose the following definition of Artificial Intelligence which compliments Wang's definition: "Artificial Intelligence is a fully controlled agent with a capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources."

References

Armstrong, S., Sandberg, A., and Bostrom, N. 2012. Thinking inside the box: Controlling and using an oracle ai. *Journal of Consciousness Studies*.

Babcock, J., Kramár, J., and Yampolskiy, R. 2016. The AGI containment problem. In *International Conference on Artificial General Intelligence*. Springer.

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., Maas, M. M., Miller, J. D., Salmela, M., Sandberg, A., Sotala, K., Torres, P., Turchin, A., and Yampolskiy, R. V. 2019. Long-term trajectories of human civilization. *foresight* 21(1):53–83.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Elamrani, A. and Yampolskiy, R. 2019. Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies* 26(5-6):35–64.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. 2017. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Legg, S. and Hutter, M. 2007. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines* 17:391–444.

Majot, A. and Yampolskiy, R. 2014. AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure. In *IEEE International Symposium on Ethics in Science, Technology and Engineering*. IEEE.

Muehlhauser, L. and Bostrom, N. 2014. Why we need friendly AI. *Think* 13(36):41–47.

Scott, P. J. and Yampolskiy, R. 2019. Classification Schemas for Artificial Intelligence Failures. arXiv:1907.07771 [cs.CY]. https://arxiv.org/abs/1907.07771.

Soares, N., Fallenstein, B., Armstrong, S., and Yudkowsky, E. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Trazzi, M. and Yampolskiy, R. 2018. Building safer AGI by introducing artificial stupidity. arXiv:1808.03644 [cs.AI]. https://arxiv.org/abs/1808.03644.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Yampolskiy, R. and Fox, J. 2012. Safety Engineering for Artificial General Intelligence. *Topoi* 32:217–226.

Yampolskiy, R. 2012. Leakproofing Singularity-Artificial Intelligence Confinement Problem. *Minds and Machines* 22(4):29–324.

Yampolskiy, R. 2013. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and Theory of Artificial Intelligence*, 389–396. Berlin Heidelberg: Springer.

Yampolskiy, R. 2015a. *Artificial superintelligence: a futuristic approach*. Chapman and Hall CRC.

Yampolskiy, R. 2015b. The space of possible mind designs. In *International Conference on Artificial General Intelligence*. Springer.

Yampolskiy, R. 2016. On the origin of synthetic life: attribution of output to a particular algorithm. *Physica Scripta* 92(1):013002.

Yampolskiy, R. 2018a. Artificial Consciousness: An Illusionary Solution to the Hard Problem. *Reti, saperi, linguaggi* 2:287–318.

Yampolskiy, R. 2018b. *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC.

Yampolskiy, R. 2019a. Predicting future AI failures from historic examples. *Foresight* 21(1):138–152.

Yampolskiy, R. 2019b. Unexplainability and Incomprehensibility of Artificial Intelligence. arXiv:1907.03869 [cs.CY]. https://arxiv.org/abs/1907.03869.

Yampolskiy, R. 2019c. Unpredictability of AI. arXiv:1905.13053 [cs.AI]. https://arxiv.org/abs/1905.13053.

Yudkowsky, E. 2011. Complex Value Systems in Friendly AI. In Schmidhuber, J., Tórisson, K., and Looks, M., eds., *Artificial General Intelligence*. Berlin Heidelberg: Springer. 388–393.

Part III

Target author's response to the commentaries in Part II

Journal of Artificial General Intelligence 11(2) 73-86, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-11-17
Accepted 2020-01-04

On Defining Artificial Intelligence —Author's Response to Commentaries

Pei Wang PEI.WANG@TEMPLE.EDU
Department of Computer and Information Sciences
Temple University
Philadelphia, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

I thank the commentators for the valuable time they put into studying and responding to the target article (Wang, 2019). They provide a wide range of perspectives on the topic representing the major positions in the field on this issue. In this response I will (1) address the issues raised about the content of my working definition, (2) discuss the overall evaluations of the definition, and (3) comment on the definitions proposed by the commentators. In the following, all references to specific sections are for those of the target article (Wang, 2019).

1. Content of My Definition

In my working definition "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources," the key points are "adaptation" and "the assumption of insufficient knowledge and responses" (AIKR), which put "intelligence" in a specific position with respect to the other concepts, such as "human intelligence," "artificial/computer intelligence," "information processing," and "problem solving."

1.1 Adaptation

Though many people consider it natural to treat intelligence as a form of adaptation, there are concerns that this requirement will make the range of intelligence too wide **(Berkeley, 2020; Rojas, 2020)** or too narrow **(Crosby and Shevlin, 2020; Laird, 2020)**. As clarified in Section 3.2, here adaptation "refers to the mechanism for a system to summarize its past experience to predict the future situations accordingly, and to allocate its bounded resources to meet the unbounded demands," so neither ELIZA nor plants qualify.

Crosby and Shevlin (2020) worried that "it risks leaving out two types of intelligent systems that we term 'Resilient Experts' and 'Fragile Geniuses'." They describe the latter as "a system that struggles with uncertainty and insufficiency, but which (intuitively) constitutes an instance of intelligence by virtue of specializing towards some particularly impressive or complex goal. ... They are wholly dependent on the cooperation of the external environment for their continued thriving and do not adapt well under AIKR conditions." As explained in Section 3.2, "adaptation refers to the attempt or effort, not the consequence," and a system with general-purpose adaptation ability will be specialized by its environment in its problem-solving skills, so this type of system is

still classified as intelligent by my definition, and I agree with their conclusion that "Most of us are Fragile Geniuses."

Since by my definition "intelligent" does not mean "successfully adapted to the environment," the "Resilient Expert" that "has rich stores of knowledge and multiple redundant mechanisms for solving any problems it encounters" so it "simply does not encounter insufficiency or uncertainty" **(Crosby and Shevlin, 2020)** are not intelligent, or at least not showing their intelligence in such a period when their performance remains unchanged, no matter how good they are in problem solving. As observed by **Laird (2020)**, according to my definition "non-learning systems, such as Chinook, Deep Blue, and Watson are not intelligent."

This leads to a central issue of this discussion: the relationship of *intelligence* and *computation*. As explained in Section 3.1, "In computer science, 'computation' does not mean whatever a computer does, but is accurately defined as a finite and repeatable process that carries out a predetermined algorithm to realize a function that maps input data to output data." **Bach (2020)** challenges this specification, though it is not my personal opinion, but how "computation" is defined by a Turing machine in textbooks on computability theory. This definition leaves no room for adaptation, as the input-output mapping carried out by a Turing machine is accurately repeatable. Such a machine always starts at the same initial state, so has no memory about what has happened in its previous runs from the initial state to the final states, nor can it learn from these runs.

It is in this sense that I contrast computation and intelligence as different ways to use a computer. It sounds contradictory to say that a *computer* can do things beyond *computation*, but any adaptive system implemented in a computer is already doing that, as far as its "input and output" is taken in the ordinary sense, that is, as problems and solutions, respectively. Of course, if the whole history of such a system is under consideration, it is still a Turing machine (Wang, 2007). For a system like NARS (Wang, 2006b), whether it is doing "computation" (or equivalently, whether it can be considered as a Turing machine or a mapping/function between its input and output) is completely determined by whether the scope of consideration is its individual inference steps, its problem solving processes, or its "life cycles" defined by its memory initialization events.

According to this analysis, intelligence is not a type of computation, but is different from it. Computation is the preferred way to use a computer if the system has sufficient knowledge and resources with respect to the problems to be solved, otherwise intelligence is the preferred way. This is why I disagree with the conclusion of **Rapaport (2020)** that AI should be based on computability theory, and why I actually do not accept Marr's conceptualization of the problem solving process in AI[1] (mentioned by **Fox (2020)**). To me, Marr accurately specified the procedure of *computational* (*algorithmic*) problem solving, but it lacks the key features of intelligence like adaptivity and flexibility.

In Section 3.2, I explained that the type of *adaptation* related to intelligence "happens within the lifetime of a single system . . . Therefore it is different from the adaptation realized via evolution in a species." Consequently, it is consistent with the proposal of **Baldassarre and Granato (2020)** "not only searches information and knowledge in the environment but it also actively builds it internally."

1. According to Marr, the problem-solving procedure consists of works on three levels: (1) defining the problem as computation, (2) designing an algorithm to carry out the computation, and (3) implementing the algorithm in a computer system (Marr, 1982).

1.2 AIKR

AIKR is undoubtedly the most controversial component of my definition of intelligence and AI. Adding it into the definition is criticized both as making the definition trivially inclusive **(Legg, 2020)** and unreasonably exclusive **(Laird, 2020)**.

Lindes (2020) suggests that "It would be better to talk about 'limits' on knowledge and resources, since 'insufficient' can only be defined relative to some task, some environment, and some performance measure." This is exactly why I choose this word. As explained in Section 4.1, "insufficient" is more restrictive than "bounded" or "limited," and is indeed with respect to the tasks the system is dealing with. Therefore, even for a system designed under AIKR, its intelligence may not show when the tasks are simple and routine.

Bach (2020) writes that "this definition does not depend on the agent itself, but on its environment, and relies on shortcomings rather than capabilities of the agent. Why would an intelligent agent that is offered unbounded resources not make use of them? Why should an agent with sufficient knowledge be considered less intelligent?" AIKR should be understood as about the system's normal *working conditions*, rather than its accidental *status quo*. As described in Section 3.2, "to acknowledge the finite nature means the system should manage its own resources, rather than merely spending them. ... Being open to new tasks means to make no restriction on the content of a task, as long as it is expressed in an acceptable form. ... For the system to live and work in real-time means that new tasks of various types may show up at any moment, rather than come only when the system is idly waiting for them, ... every task has a response time restriction."

For a specific task, it is absolutely possible that the system already has sufficient knowledge and resources, so neither learning nor creation is necessary, but that is not where its intelligence is demanded. For a difficult problem (here "difficult" is actually defined by the lack of knowledge and resources, otherwise the answer is either known or can be easily found), more knowledge and resources will indeed improve the quality of the solution, but it has nothing to do with the system's intelligence, i.e., its principles and mechanisms. Therefore, "designed under AIKR" means the system *is able to work in this situation*, rather than *happens to have a shortage*. AIKR is not a shortcoming, but a feature, even a strength, as it is exactly where the traditional models become inapplicable, as their required knowledge and resources are unavailable. Those models can be very powerful in the situations where their requirements are met, but cannot survive outside at all.

Legg (2020) writes that "Clearly any real system will have resource and knowledge limitations. In which case, why do we need to make this a part of the definition? Simply create your system and let's see how capable it is! By bringing in this additional aspect we are mixing together what a system is capable of doing, with how it goes about achieving this." The problem is that "what a system is capable of doing" depends on the *working conditions* of the system, and *theoretical* models tend to neglect the practical limitations. To design a space shuttle and to design an airplane should not be taken as the same task, mainly because these two types of "flying machines" have very different working conditions, even though both need to fly far and fast, as well as have other common features. I do not think there can be a non-trivial objective for the design of all flying machines irrespective to their flying conditions.

Legg (2020) further adds: "When I have made this point in the past, some people then ask why I am interested in things like AIXI. I see AIXI in the same way as I view Turing machines: as an abstract model and allows a certain kind of theoretical analysis, not as a blueprint for actually building a real system." A theory of AI does not have to be a blueprint, but still need to provide

some guidance to the building of actual systems. At least for systems like NARS, AIXI is mostly irrelevant.[2] Beside resource restriction, the actions of NARS cannot be directly evaluated in terms of an expected utility, but only according to the concrete goals described using the concepts generalized from the system's experience. The system's environment cannot be considered as a probability distribution over Turing machines, as the system has to express its experience in a much higher level of abstraction than streams of input symbols or signals. The analogy of Turing machine is not acceptable here, as the concept of *computation* has no requirement on resources (except the processing time should be finite), nor is the need of (lossy) compression/abstraction/generalization of experience, but the concept of *intelligence* requires these factors to be included directly or by implication, and omitting them will completely change the nature of the problem.

1.3 Intelligence in concept hierarchy

Some commentaries are mainly about the position I place intelligence within the whole conceptual system of science, as the meaning of a concept is not only specified by its boundary, but also by its relations with other concepts.

Lindes (2020) and **Rojas (2020)** question my usage of "information" and "information-processing system." Though these concepts have their own controversies, in this discussion "information-processing system" is simply used as the superordinate concept of "intelligent system" and "computational system." This usage does not touch issues like "where do we draw the line between information and knowledge?" **(Rojas, 2020)** at all, but just provides a common platform on which the comparison between intelligence and computation can be carried out. As explained in Section 3.2, "information-processing system" is used "to include all computer systems and robotic devices, as well as many animals, though it will not include everything, such as rocks and rivers," so it will also not include the Earth. In this context, information and its processing are used as *methodological*, rather than *ontological* concepts, in the sense that the important question about them is not whether they exist, but what benefit they can provide in the description of objects and events. I use these concepts to raise the description of a system to an abstract level, so systems with various substances (such as machines and animals) can be compared with respect to their structures and functions, without touching the differences in how the objects are composed and how the events are carried out.

Rojas (2020) and **Yampolskiy (2020)** stress the differences among *general intelligence*, *human intelligence*, and *machine intelligence*. I surely agree that they are separate concepts, though I disagree with the conclusion of **Rojas (2020)** that "Human intelligence is based on pattern recognition, intuition and filtering of unnecessary details. Computer intelligence is based on fast electronics and optimal algorithms." To me, the basic principle and mechanism of human intelligence and computer intelligence (AI) should be basically the same, which is what the general notion of intelligence is, with human intelligence and computer intelligence, as well as some others, as special forms, which are also restricted and shaped by the substance of the implementation (biological and electrical, respectively).

Laird (2020) proposes using explicit modifiers to "intelligence," which will indeed reduce the confusion to a certain extent. Beside the above "human vs. computer" distinction, I also used this approach in Section 2.2 when summarizing the attempts of defining AI. However, the underlying

2. AIXI is a model of intelligence described in (Hutter, 2005).

problem is still there: if they are so different from each other, why do they all have "intelligence" in it? What do these concepts have in common?

As **Mikolov (2020)** suggests, "although the problem of defining AI is a very difficult one, it might be considerably easier to define what AI is not." To me, in this context "intelligent" has two groups of antonyms, one includes "dull," "stupid," "foolish," etc., while the other includes "instinctive," "innate," "mechanical," "computational," etc. While the former has a negative flavor associated with it, the latter just indicates a problem-solving mechanism (or mode) that is fundamentally different from intelligence.

When the concept of *intelligence* was introduced to discuss human intellectual capabilities, two factors were merged together, where one is the concrete capabilities of solving specific problems, and the other is the meta-level capability of acquiring and improving these specific capabilities. As analyzed in (Wang, Liu, and Dougherty, 2018), that treatment is acceptable in psychology, as the capabilities of newborn humans are similar enough for the problem-level and meta-level capabilities to be considered as roughly correlated. However, this ambiguity is unjustifiable in AI, as the problem-solving capabilities and the learning capabilities of AI systems are not correlated at all. To call both "intelligence" is a major cause of the current confusion.

My definition reserves "intelligence" for the unified, meta-level, domain-independent capability, and I use "skill" for the diverse, problem-level, domain-dependent capabilities. Consequently, systems that cannot learn are not intelligent at all, though they can be very skillful in problem solving, as in what **Crosby and Shevlin (2020)** called "Resilient Experts" or the famous AI systems mentioned by **Laird (2020)**. On the other hand, the intelligent systems are not necessarily skillful, especially in novel or radically changing environments.

This distinction between intelligence and skills is related to the one between domain-general and domain-specific cognition **(Baldassarre and Granato, 2020)**, as well as to that between Artificial General Intelligence (AGI) and AI. To me, AGI is nothing but AI in its initial and ultimate sense, but not the current mainstream AI, because most people are working on skills, not intelligence, using my terminology. I am not saying skills are not valuable, but that they are fundamentally different from intelligence. General-purpose capability cannot replace the value of special-purpose capabilities, so "AGI is not proposed as a competing tool to any AI tool developed before, by providing better results, but as a tool that can be used when no other tool can, because the problem is unknown in advance" (Wang and Goertzel, 2007).

This level distinction is another reason why my definition of intelligence is different from those focusing on problem-solving capabilities. Of course, the meta-level capability will eventually show its effects in problem-level capabilities, but it is not the reason to deny their differences. If people still think it is more natural to use "intelligence" for problem-solving capabilities, I do not mind to use other words for the meta-level capability, such as "cognition." Since I also have a background in cognitive science, I agree with **Rapaport (2020)** on seeing "cognition" as basically the same as "intelligence" in this context, and with **Fox (2020)** on the close relationship between the ultimate goals of cognitive science and AI. However, even if we change the label, it still does not change the fact that many key "AI problems" are actually at the meta-level, not the problem-level. As I argued in Section 1.2, the choice of words is a secondary problem.

Baldassarre and Granato (2020) argue "that intelligent systems should be based on sub-symbolic representations and parallel distributed processing, as those of neural networks, rather than on symbolic representations and logic inference." In Section 4.2.1, I briefly explained that I do not take this approach due to consideration of *generality* and *necessity*. As stated in Section 3.3,

NARS uses an *experience-grounded* semantics. According to it, the meaning of a term in NARS is determined by its experienced relations with other terms within the system, rather than by the external object or event it refers to, as in "symbolic AI" systems. In this way, the representation in NARS becomes *semi-distributed*, and the traditional "symbolic vs. sub-symbolic" distinction cannot be made anymore. NARS actually shares many properties with neural networks, though does not explicitly simulate the brain structure (Wang, 2006a).

Bach (2020) observes that "Pei Wang does indeed believe that human intelligence is close to the limit of that of any possible intelligent system, although he accepts that the capacity for self modification is an important part of intelligence." This is again related to the meta-level vs. problem-level distinction. At the level of concrete problem-solving processes, intelligence implies self-modification, and it is perfectly possible, and even inevitable, for AI to outperform human beings in solving more and more problems, including to find solutions we never think of and have difficulty to fully understand. On the contrary, at the meta-level intelligence is nothing but an advanced form of adaptation. Though AI systems can be larger and faster than a normal human mind, there is no evidence that they can self-improve beyond the concept of adaptation into a form of existence outside our comprehension completely. This is why my definition of intelligence leaves no room for notions like "superintelligence" and "singularity" (Wang, Liu, and Dougherty, 2018).

2. Overall Evaluation of My Definition

Now let me zoom out from the specific points to the evaluation of my definition as a whole. As no one has challenged the validity of the four criteria originally from Carnap, I will continue to use them to organize the comments and responses.

2.1 Similarity to the explicandum

This is predictably the most debatable point, as my definition is obviously different from the common definitions in AI textbooks and surveys. As recognized by **Berkeley (2020)** and **Laird (2020)**, by my definition many systems that are currently considered as AI will be judged as not intelligent at all.

This is exactly why in the target article I clearly distinguished a "working definition" from a "dictionary definition." When reviewing a submission for an AI conference or journal, I will not reject it as irrelevant because its definition of AI is different from mine, and in such a case I obey the dictionary definition. On the other hand, I explicitly announce that my own research is targeted at an objective that is very different from most of the goals pursued by the other AI researchers. It is not always easy to joggle two incompatible definitions, though not impossible.

Actually, the same incompatibility happens between the definitions of AI currently hold by the mainstream AI community and what the public considers as "AI," and AI researchers often complain that the public has a high and unrealistic expectation of AI. However, the same observation can also be interpreted as that the AI community has trivialized the concept of intelligence to something easier and more feasible. In a sense, my working definition is arguably closer to the public view of what intelligence should be about, by directly associating the concept to adaptivity, flexibility, originality, etc.

Stone (2020) writes "From the highest level perspective, I agree with Wang's exposition of the values of specifying one's working definition, and commend him for acknowledging on more than one occasion that there is room for different definitions. But despite this acknowledgement, I note

that on more than one occasion he seems to argue for the need to converge on a single definition or the superiority of his own definition, neither of which I endorse." His statements about my position is accurate, though I do not feel any contradiction. I consider the field of AI still in a pre-paradigmatic state, so it is necessary to tolerate different opinions and to encourage new ideas (including working definitions). However, it does not mean that every idea is equally good, or that we should not make the effort to compare the definitions and *attempt* (not *force*) a convergence. As **Bach (2020)** puts it, "I think this implies that Artificial Intelligence research has to concern itself with studying the nature of intelligence. If it succeeds and identifies its subject, a lot of the working definitions will either disappear or turn out to relate to aspects of the same subject, and be replaced by a functional one." Even if it turns out that a single working definition of intelligence is impossible, we can still expect to end up with a small number, with a relatively clear relationship among them. I surely consider my own working definition to be superior, and I assume everyone in the field also considers one's own working definition the best, otherwise why does he or she still hold it?

Stone (2020) states "Personally, I hold strongly to the 'big tent' view of AI that allows, and even encourages, multiple perspectives and agendas, and thus working definitions, to co-exist within the same field." To me, this is exactly where the "identity of AI" problem comes from. "To let different approaches coexist and compete in solving a problem" is not the same as "to use the same name for many fundamentally different problems." In Section 2.2.6, I argued that the researchers in the current AI field are not climbing the same summit. Though it does not prevent us from respecting each other and learning from each other, even cooperating with each other, to put these climbers with different destinations in the same tent will make it hard to draw any non-trivial conclusion about them. If every conclusion is only about some of them, what is the benefit of clustering them together in this tent? There is surely a historical reason, but that does not imply a necessity for the present time.

2.2 Exactness

Chollet (2020) concludes that "Overall, Wang's definition, while grounded in a very reasonable and even wise vision of intelligence, falls short of its own goals of 'drawing a sharp boundary' and 'being fruitful', due to insufficient formalism and excessive reliance on implicit semantics."

In Section 1.3.2, I stated that "formal definitions are preferred, as they are generally more accurate and less ambiguous," and then added that "since the concept of intelligence has empirical content, its definition cannot be completely formal" and "the existence of different interpretations may undermine the exactness of the definition."

For example, in Section 2.2 a simple formal model of "agent" is introduced, consisting of its input signals, internal states, and output actions, so as to separate different abstractions of human intelligence. This formal model is more exact than talking about "human-like" AI without specifying where the likeness is. **Bach (2020)** comments "percepts and actions cannot be readily treated as an interface to the environment. Instead, percepts and actions are themselves representational states. An understanding of perception and action will generally not be independent of the model of intelligence of the agent itself, hence making the comparison between different approaches in this framing difficult or even impossible." As far as the comment itself is concerned, I basically agree, and I went even further to challenge the separation between perception and action (Wang and Hammer, 2018). However, the purpose of this formal model is merely to disambiguate

different types of "human-like," rather than to serve as a full model of intelligence, therefore rough treatments are taken to make the description simple. This is the problem of formal models, where sharp lines often oversimplify the problem, as in my previous criticism of AIXI.

Chollet (2020) notices "a large jump in Wang's argument between the vagueness of his working definition and the high specificity of his work on NARS and NAL" which indeed exists. NARS can be taken as a formalization of my working definition of intelligence, though I do not use it, or a simplified version of it, as a working definition, for several reasons:

- Even after simplification, such a formalization will still be too complicated to serve as a working definition.
- Interpretations of the symbols may decrease or damage the exactness of the formal definition.
- Though NARS faithfully realizes my definition, it is not necessarily the only possible realization.

When we stress the importance of a feature (such as *exactness*), it does not mean that we will pursue it and ignore the others. When deciding where to go, it is not easy to exactly point to the right direction. In such a situation, I would rather vaguely point to the direction that felt right to me than exactly point to a direction that has recognizable flaws.

To desire an exact definition of intelligence does not contradict with the acknowledgment that intelligence is a matter of degree. On this matter I agree with **Rojas (2020)** and **Rosa (2020)**. My definition makes the intelligence of some systems comparable, as described in Section 4.4: "one system can be more intelligent than another by being able to acquire knowledge in more forms (e.g., additional sensorimotor channels), to reorganize its beliefs and skills in more complicated ways (e.g., more recognizable patterns), or to adapt more efficiently (e.g., faster responses)." The quantile of a system among comparable ones can be taken as a rough measurement of its intelligence, that is, a value 0.8 means "more intelligent than 80% of the comparable systems, and less intelligent than the other 20%". Here the crucial point is to measure the meta-level learning capability, rather than the concrete problem-solving capability (Wang, Liu, and Dougherty, 2018). In this respect, it is similar to the idea of **Rosa (2020)** that "the key metric is how fast it adapts and learns to solve novel tasks."

2.3 Fruitfulness

I full agree with **Chollet (2020)** that "Ultimately, practical impact in the real world is the scale by which the value of a working definition of AI will be weighted." However, since all AI projects are far from their ultimate destinations (except those who trivialize the concept of intelligence and claim AI has been fully realized), we must compare the *present* results, maybe plus the potential results that have a high plausibility in the near future.

The most direct result of my working definition is NARS, which is described briefly at Section 3.3 and in detail in my other publications. **Chollet (2020)** comments that "one may feel that the arguments behind Wang's definition were retrospectively conceived to justify the work on NARS and NAL," which is partially correct, as the definition has been formed and confirmed *during* NARS design and development, rather than completely *before* or *after* it. Besides *justifying* the engineering work, the more important role of this working definition is *guiding* the work, as explained in Section 3.3.

NARS uniformly realizes many cognitive functions (listed in Section 3.3), though in the target article I did not mention any practical application of NARS, or its performance on the common AI tasks. Several commentators **(Bach, 2020; Crosby and Shevlin, 2020; Fox, 2020; Mikolov, 2020)** raise this as an issue on the value of the working definition.

The lack of practical result has several reasons:

- NARS has not been finished yet. As a system whose components are closely coupled with one another, even a mostly-finished version is hard to use in practical situations. Though there have been some experiments in which some components of NARS (such as a subset of its inference rules) are used for practical purposes, they do not qualify as applications of NARS.

- As my definition and theory focus on meta-level, an out-of-box NARS, even after it is "fully built" (in a certain sense), still has little skill when facing practical problems, just like a newborn baby. Its intelligence is in its *potential*, rather than in its *current* abilities. To turn the former into the latter, an *education* process is needed, which is fundamentally different from the *training* processes in the current machine learning systems, and we still have not all the details worked out.

In recent years our team has been cooperating with a team in Cisco to develop applications of NARS in various domains. The preliminary result in the "smart city" domain is reported in (Hammer et al., 2019), and the functionality is being integrated into the products of Cisco. As we have other on-going application-oriented projects, there is reason to expect more practical techniques coming out of this research in the near future.

I agree with **Fox (2020)** that "Medicine draws on a vast diversity of knowledge and human skills and requires many different forms of intelligence." We did some experiments in that domain (Wang and Awan, 2011), and have been working on it in recent years.

As explained previously, NARS is not designed to compete with the existing AI techniques on the problems they are designed to solve, but to solve a (meta-level) problem "how to adapt under AIKR," which has got little attention in the AI community. Even so, NARS is still related to the other AI theories and techniques here or there, and there are publications to compare NARS with the other techniques. For example, NARS has been compared with neural networks (Wang, 2006a; Wang and Li, 2016) and reinforcement learning (Wang and Hammer, 2015), which are related to the comment of **Crosby and Shevlin (2020)** on the relation between Deep Reinforcement Learning and Artificial Intelligence.

2.4 Simplicity

Stone (2020) points out that my definition is not as simple as it looks, "Rather, to fully understand it requires reading its explanation throughout the 2.5 pages of Section of 3.2." This is correct, and for a complicated and multifaceted concept like intelligence, it is hard to get a simple but nontrivial definition. To fully understand my position and my reasons to take this position, it is necessary to read my technical writings, even the source code of NARS. It is just like what **Rapaport (2020)** concludes, "As a consequence, one-sentence definitions such as any of those under discussion are really only acceptable for quick overviews or dictionaries. To really understand a subject, one needs at least an encyclopedia article, a textbook, or a research program."

The solution, I believe, is a compromise among the requirements, including a balance between simplicity and the other requirements for a working definition, as discussed in Section 1.3.5. I have tried my best, and am open to suggestions on how to improve it, or to replace it by a better one.

3. Other definitions in the commentaries

About half of the commentators propose their own definitions of intelligence. They are listed below and followed by my brief comments, mainly to highlight their differences from mine.

Bach: *[Intelligence is] the ability to deal with complexity by making models, usually in the service of a complex control task (such as the persistent existence of a complex agent in an entropic universe)* **(Bach, 2020)**.

—I assume that in order to exist in an entropic universe, adaptation becomes necessary, and if "complex" is interpreted as similar to AIKR, I mostly agree. What counts as "making models" may be an issue. The beliefs in NARS do not form a model of the objective world, but a summary of the system's experience, which is fundamentally subjective.

Baldassarre and Granato: *[I]ntelligence is the capacity of an agent to use computation, intended as the capacity to link perception to action in multiple possibly sophisticated ways, to increase biological fitness or to accomplish goals. . . . General-domain intelligence is the capacity of goal-directed agents to flexibly accomplish novel goals in novel conditions/domains by building the knowledge they lack through the manipulation of internal representations and by actively seeking such knowledge in the external environment* **(Baldassarre and Granato, 2020)**.

—I think their "intelligence" *tout court* is basically what I call "skills" previously, while their "general-domain intelligence" is closer to my definition. The demand of being "goal-directed" is trivial, as every computer program may be interpreted as goal-oriented. To me, goal-oriented activity is a feature of information processing, either intelligent or not. I agree with the requirements implied by "flexibly" and "novel," though feel that they need to be further specified—they roughly correspond to the *open* component of AIKR.

Laird: *Intelligence is a measure of the optimality of behavior (actions) relative to an agent's available knowledge and its tasks, where a task consists of goals embedded in an environment* **(Laird, 2020)**.

—I agree with the first half, and would add "resources restrictions" to it, especially the demand of real-time responses. It is unclear whether the goals are predetermined, or whether the environment may change radically. To me, these factors are what really matter in the definition of intelligence, since to achieve a constant set of goals optimally in a static environment needs little intelligence.

Lindes: *Intelligence is the ability of an agent, whether human, animal, artificial, or something else, to act in its environment in real time, using its limited knowledge, memory, computational power, and perception and action capabilities, choosing actions at each moment that move it toward its current goals, and to adapt over time by improving this ability to act* **(Lindes, 2020)**.

—Mostly agree. Compared to mine, a major difference is that it does not explicitly mention that the goals may be beyond the system's current capability and be inconsistent with one

another, so cannot be achieved together. Such a situation is implied by the *open* component of AIKR.

Rapaport: *AI is a branch of computer science (CS), which is the scientific study of what problems can be solved, what tasks can be accomplished, and what features of the world can be understood computationally (i.e., using the language of Turing machines), and then to provide algorithms to show how this can be done efficiently, practically, physically, and ethically* **(Rapaport, 2020)**.

—This definition effectively takes AI problems as a subset of CS problems. As explained previously, I believe AI problems are conceptually beyond the scope of computability theory, though AI systems are still implementable in computers.

Rosa: *[W]e defined intelligence as a "problem-solving tool that searches for solutions to problems in dynamic, complex and uncertain environments"* **(Rosa, 2020)**.

—According to the interpretation of Rosa, their working environment obeys AIKR. In that case our difference is relatively small, though I would rather not call intelligence a "problem-solving tool," or use "search for solutions" to describe its basic function, given the previous discussion on the distinction between problem-level and meta-level.

Stone: *[From the 2016 report of the One Hundred Year Study on AI] Artificial Intelligence (AI) is a science and a set of computational technologies that are by inspired by—but typically operate quite differently from—the ways people use their nervous systems and bodies to sense, learn, reason, and take action* **(Stone, 2020)**.

—This is a good dictionary definition, but not a good working definition, as it leaves too much space for the interpretation of the concepts involved, and does not provide much guidance for the following research. AI has indeed got its inspirations from the human brain/mind complex, though these inspirations often point to different directions when a concrete design decision is made, as analyzed in Section 2.2.6.

Sutton: *[From John McCarthy] Intelligence is the computational part of the ability to achieve goals in the world* **(Sutton, 2020)**.

—This definition is too broad. As I said previously, every computer program has the ability of achieving certain goal in the world. If they are all considered intelligent, this label has no meaning.

Yampolskiy: *Artificial Intelligence is a fully controlled agent with a capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources* **(Yampolskiy, 2020)**.

—Here the suggested addition to my definition is that an AI system must be *fully controlled*, so as to guarantee its safety. Though I fully agree we should make AI safe, I do not think it can be achieved by using such a working definition to exclude the uncontrollable systems from the category. Furthermore, "fully controlled" may be interpreted very differently. For an adaptive system like NARS, it cannot be controlled by restricting its initial design alone, while it can be fully controlled by restricting its initial design as well as its lifelong experience, though the second part is hard to realize in practice. Is such a system "fully controlled"?

4. Summary

I hope this JAGI special issue is the beginning, not the ending, of this discussion. We surely do not want to spend all our time on debating definitions, and the final judge of this competition is time, but nevertheless I believe the current situation is that the problem is getting too little attention, not too much. This is especially true when many other debates in AI can be traced back to the different understandings of intelligence.

It is just normal for every researcher to believe they have the best idea, so we cannot expect some consensus to be achieved soon, but at least everyone should make his/her research objective relatively clear, which will reveal its preconditions and consequences, and reduce miscommunications.

Finally, I want to thank the editors for their tremendous efforts in organizing such a broad and deep discussion.

References

Bach, J. 2020. When Artificial Intelligence Becomes General Enough to Understand Itself. Commentary on Pei Wang's Paper "On Defining Artificial Intelligence". *Journal of Artificial General Intelligence* 11(2):15–18.

Baldassarre, G. and Granato, G. 2020. Goal-Directed Manipulation of Internal Representations Is the Core of General-Domain Intelligence. *Journal of Artificial General Intelligence* 11(2):19–23.

Berkeley, I. 2020. AI: A Crowd-Sourced Criterion. A Commentary on Pei Wang's Paper "On Defining Artificial Intelligence". *Journal of Artificial General Intelligence* 11(2):24–26.

Chollet, F. 2020. A Definition of Intelligence for the Real World? *Journal of Artificial General Intelligence* 11(2):27–30.

Crosby, M. and Shevlin, H. 2020. Defining Artificial Intelligence: Resilient Experts, Fragile Geniuses, and the Potential of Deep Reinforcement Learning. *Journal of Artificial General Intelligence* 11(2):31–34.

Fox, J. 2020. Towards a Canonical Theory of General Intelligence. *Journal of Artificial General Intelligence* 11(2):35–40.

Hammer, P., Lofthouse, T., Fenoglio, E., and Latapie, H. 2019. A reasoning based model for anomaly detection in the SmartCity domain, NARS Workshop at AGI-19, Shenzhen, China. `https://cis.temple.edu/tagit/events/papers/Hammer.pdf`.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer.

Laird, J. 2020. Intelligence, Knowledge & Human-like Intelligence. *Journal of Artificial General Intelligence* 11(2):41–44.

Legg, S. 2020. A Review of "On Defining Artificial Intelligence". *Journal of Artificial General Intelligence* 11(2):45–46.

Lindes, P. 2020. Intelligence and Agency. *Journal of Artificial General Intelligence* 11(2):47–49.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman & Co.

Mikolov, T. 2020. Why Is Defining Artificial Intelligence Important? *Journal of Artificial General Intelligence* 11(2):50–51.

Rapaport, W. J. 2020. What Is Artificial Intelligence? *Journal of Artificial General Intelligence* 11(2):52–56.

Rojas, R. 2020. On Pei Wang's Definition of Artificial Intelligence. *Journal of Artificial General Intelligence* 11(2):57–59.

Rosa, M. 2020. On Defining Artificial Intelligence—Commentary. *Journal of Artificial General Intelligence* 11(2):60–62.

Stone, P. 2020. A Broader, More Inclusive Definition of AI. *Journal of Artificial General Intelligence* 11(2):63–65.

Sutton, R. S. 2020. John McCarthy's Definition of Intelligence. *Journal of Artificial General Intelligence* 11(2):66–67.

Wang, P. and Awan, S. 2011. Reasoning in Non-Axiomatic Logic: A case study in medical diagnosis. In Schmidhuber, J., Thórisson, K. R., and Looks, M., eds., *Proceedings of the Fourth Conference on Artificial General Intelligence*. Springer. 297–302.

Wang, P. and Goertzel, B. 2007. Introduction: Aspects of artificial general intelligence. In Goertzel, B. and Wang, P., eds., *Advance of Artificial General Intelligence*. Amsterdam: IOS Press. 1–16.

Wang, P. and Hammer, P. 2015. Assumptions of decision-making models in AGI. In Bieger, J., Goertzel, B., and Potapov, A., eds., *Proceedings of the Eighth Conference on Artificial General Intelligence*. Springer. 197–207.

Wang, P. and Hammer, P. 2018. Perception from an AGI perspective. In Iklé, M., Franz, A., Rzepka, R., and Goertzel, B., eds., *Proceedings of the Eleventh Conference on Artificial General Intelligence*. Springer. 259–269.

Wang, P. and Li, X. 2016. Different conceptions of learning: Function approximation vs. self-organization. In Steunebrink, B., Wang, P., and Goertzel, B., eds., *Proceedings of the Ninth Conference on Artificial General Intelligence*. Springer. 140–149.

Wang, P., Liu, K., and Dougherty, Q. 2018. Conceptions of artificial intelligence and singularity. *Information* 9(4):79.

Wang, P. 2006a. Artificial general intelligence and classical neural network. In Zhang, Y.-Q. and Lin, T. Y., eds., *Proceedings of the IEEE International Conference on Granular Computing*. Atlanta, Georgia: IEEE.

Wang, P. 2006b. *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.

Wang, P. 2007. Three fundamental misconceptions of artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence* 19(3):249–268.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.

Yampolskiy, R. V. 2020. On Defining Differences Between Intelligence and Artificial Intelligence. *Journal of Artificial General Intelligence* 11(2):68–70.

Part IV

Other invited peer commentaries addressing
the definition of artificial intelligence

Journal of Artificial General Intelligence 11(2) 89-90, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-17
Accepted 2019-11-07

What Is AI?

Roger Schank ROGER@SOCRATICARTS.COM
Socratic Arts Inc.
John Evans Professor Emeritus, Northwestern University
Evanston, Illinois, USA

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

AI is a poorly chosen term. The term confuses everyone, including those who work in AI. I made the mistake of inviting John Searle to visit my AI Lab at Yale 40 years ago (or so) and he heard some of my students expound what he later called the *strong AI hypothesis*. He asked if my students thought computers would actually be intelligent and many said they would. (I assume they were some of my newer students.) He never asked me because he knew that I would say the idea was absurd.

The best way for me to explain AI is through the idea of "reminding" an idea that I have been discussing for the same 40 years.

The classic example of reminding is *the steak and the haircut* story: A colleague of mine responded to my complaint about the fact that my wife couldn't cook steak as rare as I wanted it by saying that twenty years earlier, in London, he couldn't get his hair cut as short as he wanted it. While this may sound like a brain-damaged response, these two stories are identical at the right level of abstraction. They are both about asking someone to do something who, while being capable of doing it, has refused to do it, because they thought the request was too extreme. My friend had been wondering about his haircut experience for twenty years. My story reminded him of his own experience and helped him to explain to himself what had happened.

This is my key example of how intelligence works. We hear stuff, see stuff, read stuff, and our memories, quite unconsciously, come up with things to think about that are similar. We must do this because we need to recognize people, places, situations, prior thoughts and so on to help us interpret everyday experiences. Insights, new ideas, decisions about how to respond to something, all require the considerations of prior ideas, decisions, responses etc. An intelligent entity remembers its mistakes so it doesn't keep repeating them.

What does this have to do with AI? It depends upon whose AI you are talking about. It has nothing to with modern AI which is mostly about counting, pattern matching, and statistics, all of which is not even close to how the human mind operates (or a dog's mind either.)

There have always been two very different definitions of AI. Early on, AI was very focused on creating programs that can play chess. I remember those days very well. I asked then why AI was focused on chess and the answer was always about trying to show that computers could be "smart." In other words, there were these new things called computers and researchers wanted to see what cool things they could do. No one needed a computer that played chess. This was one kind of AI—building cool stuff. (John McCarthy was the key guy in this kind of AI.)

The other kind was focused on how grand masters played chess. This kind of AI (Newell, Simon, and Minsky were the leaders) was not about whether you could build cool stuff but about a new way of working on psychology—seeing how much we could learn about the human mind by trying to replicate what it can do. These two different approaches were filled by two different types of people. The first type were technical people and the second type were what would now be called cognitive scientists.

My own field at the time (*computational linguistics*, a name I changed to *natural language processing* because I didn't consider what I was working on as being part of linguistics) had the same two types of approaches. There were those who only cared about syntax (and therefore parsing sentences) because they were followers of Chomsky and really weren't concerned with meaning, understanding, or how the mind works. I didn't care about parsing diagrams or syntax. I cared about understanding language—about how people do it and how I could get computers to do it.

This was a long time ago but as the French say, the more things change the more they stay the same.

So, today we have AI that is not about people and this kind of AI has taken over the field. This is due to constant hype by the media and heavy investment from venture capitalists. The same thing happened in the mid 80's. Expert systems were being hyped then in the same way that "deep learning" is being hyped now. And the media and the VCs are all over it once again. In 1984, I ran a panel at AAAI about what we should do about the coming *AI Winter*. One could see that "Expert Systems" would soon be over because they didn't work very well. Why didn't they work well? One reason is illustrated by the Steak and the Haircut story. These "experts" never got reminded of their previous experiences. They couldn't reason from prior cases. They just had rules that had been gathered by researchers. Those rules were based on the assumption that the experts know what they know and they can tell you the rules they know. But it follows from looking at how reminding works, that people do not know what they know. They get reminded without knowing how that happens because they have very little insight into how their minds work.

I am interested in how human memory works. These days thinking about the human mind is not considered AI. I would like computers to get reminded of something they know by something that just happened. But that can't happen unless the computer can store and retrieve stories indexed by their meaning. A good plan would be to discover the method that humans unconsciously use to find things that they know when they need them.

It is unfortunate that we are not trying to index important knowledge based upon particular experiences in the computer. People like to talk about their experiences. Why? How do we get a computer to want to tell you what it has just experienced? Computers would be rather important for the advancement of knowledge and civilization if we could hear their wisdom. Of course, they would have to have some wisdom first. How do we make them wise? People get wiser every day without actually trying. Why don't computers? We need to take such questions seriously before there will be any actual AI.

But those kinds of questions are no longer asked in AI. AI is now just about counting. And while that may well be useful in certain situations, it is not intelligence and it would be good if we would stop calling it AI.

Journal of Artificial General Intelligence 11(2) 91-96, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-10-04
Accepted 2020-01-16

A Philosopher-Scientist's View of AI

Aaron Sloman HTTP://WWW.CS.BHAM.AC.UK/~AXS
School of Computer Science
University of Birmingham
Edgbaston, Birmingham, UK

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Since the "official" launch of AI in 1956, preceded by earlier mathematical and philosophical work by Turing and even earlier practical uses of automated calculators and controllers of various sorts, AI has included a wide range of activities, by scientists, engineers, and others with widely varying aims, now mostly dominated by practical, engineering aims. Some of the early work had scientific and philosophical, rather than engineering goals. My own work is of the former type, including use of AI to investigate architectural ideas about how *cognitive* functions interact with motivation, emotions and other varieties of *affect*, addressing old problems in philosophy and the sciences of mind. Some of the difficulties encountered suggest that modelling/replicating ancient mathematical and spatial reasoning abilities of humans and other intelligent species may require digital computers to be enhanced with mechanisms that combine discrete and continuous forms of computation, in ways that nobody understands at present, although sub-neural chemistry-based mechanisms with such a combination are attracting increasing attention. Regarding the recent use of the label "AGI" (Artificial General Intelligence) I have always assumed that AI should accommodate any mechanisms that work, including specialised subsystems common in robotics, so adding a "G" for "general" seems to me to be a misleading publicity gimmick.

1. Introduction: Surveys by pioneers

Anyone wishing to understand the scope and methods of AI can still benefit from the vision of some of the pioneers, not because they had a right to limit future developments, but because their work often included useful/powerful ideas that are still important. Minsky's remarkable survey originally written around 1960 (Minsky, 1963) with over 100 bibliography entries (and still downloadable from his web site[1]) included many such ideas. An important early publication recognizing implications of AI for psychology, was Miller, Galanter, and Pribram (1960). In 1969, an important, but more methodologically focused, paper on the scope and methods of AI from a philosophical standpoint was McCarthy and Hayes (1969), arguing that *logical* forms of expression are *metaphysically*, *epistemologically*, and *heuristically* adequate forms of representation for intelligent machines. Those ideas are still used by many AI researchers employing logic-based representations, sometimes in hybrid systems, e.g. combined with diagrammatic or probabilistic reasoning, challenging the heuristic adequacy of pure logic-based AI, as in (Sloman, 1971).

1. https://courses.csail.mit.edu/6.803/pdf/steps.pdf

Like many branches of pure and applied science, AI builds on earlier achievements, including designs for calculators and controllers (e.g. automated looms), as well as research in logic, philosophy, psychology, neuroscience, linguistics and social sciences. AI has always included research with both scientific and philosophical aims, although engineering aims and achievements now dominate news about AI. Research fields can also include participants focusing on very different aims, e.g. some more interested in solving old practical problems, some seeking new explanations for old phenomena, and some seeking new practical applications.

This paper focuses on relationships between AI and natural intelligence that are not always acknowledged or widely understood. As indicated above, AI has always been far more than an engineering discipline concerned with making smart machines. For example, Alan Turing, John von Neumann, John McCarthy, Herbert Simon, and Marvin Minsky were as interested in explaining natural intelligence, and, in some cases, answering philosophical questions, as in making smart new machines. I'll also try to show that there are deep explanatory gaps in current AI that generally go unnoticed, and which may require development of new forms of computation. Any attempt to *define* "Artificial Intelligence" should at least allow for the possibility that over time it can change its aims and methods and mechanisms, at least as much as physics has done since ancient attempts to understand such things as levers and planetary motion. Some of this evolution was documented in great detail in (Boden, 2006).

So attempting to *define* AI in terms of its current tools and aims at any time is seriously misguided. Despite his breadth of vision, McCarthy was disconcerted by the suggestion in Sloman and Croucher (1981) that some intelligent machines will unavoidably have emotions, as a side-effect of design requirements for intelligence with limited knowledge and resources. He thought AI systems should be prevented from having emotions, since that could reduce their reliability. In part this reflects a difference between AI as engineering and AI as science. On that occasion, McCarthy's scientific and philosophical goals were to some extent blunted by his engineering goals. Contrast the broad aims of Minsky (2006).

Debates about what should be included in AI risk being pointless, like some debates about the scope of mathematics: e.g. does mathematics (or AI!) include parts of theoretical computer science? Debates about what should be included in education for young learners are not pointless, however, because restricting diversity in education can have bad effects. Instead of stipulating boundaries it is more important that AI researchers and teachers (like all other researchers and teachers) are clear about their explanatory or practical goals, how they relate both to preceding ideas and possible future developments, and when disagreements about goals are not disagreements about facts. Although individual teachers or schools cannot cover everything relevant national educational systems should allow, and even encourage, diversity, in order not to hobble future research.

People offering services, products, courses, degrees and certificates should, of course, be clear about the scope of what they are offering, but stipulating *definitions*, especially for research fields, can restrict freedom to explore new directions and may block scientific and engineering advances, as well as constraining educational opportunities for young minds. Historical surveys may limit their scope provided they acknowledge incompleteness, as Boden does (1977; 2006).

2. Pattern recognition *vs* AI scene analysis

Sometimes disagreements about the scope of AI, or branches of AI, are based on different assumptions about natural intelligence. For example, a strand in AI since its earliest days was

pattern recognition, designing self-extending programs trained to segment recorded speech into words, phrases, sentences, etc., or 2D visual images into 2D portions with learnt labels attached, e.g. "head", "arm", "finger", "eye", in contrast with the *scene analysis* approach adopted by Clowes and others in the late 1960s, attempting to use 2D input image structures (e.g. lines, line-junctions, and 2D regions) to *derive* descriptions of 3D structures with parts and relationships, on the basis of general principles of projection, or attempting to derive semantic structures from written or spoken language input using syntactic and semantic theories, sometimes augmented with prior world knowledge. For example, a junction in a 2D image where several lines meet might be interpreted as representing a 3D vertex where several edges meet, some interpreted as convex and some concave, even if that particular configuration of lines and junctions had never previously been encountered in a "training" session (Clowes, 1971, 1973).[2] A crucial feature of such work was use of context to resolve local ambiguities—important in both language understanding and visual perception. Later research extended the ontologies used by such scene analysis systems.

The 1960s AI work in vision was partly inspired by work in linguistics, e.g. Chomsky (1965), on the relationships between syntactic structures in sentences and semantic descriptions of portions of the world. Clowes was also influenced by ideas in (Abercrombie, 1960), concerning visual learning in trainee medical researchers learning to derive descriptions of minute physiological structures from images perceived using microscopes. Gombrich (1960) also influenced AI vision researchers.[3] Proceedings of the 2nd IJCAI (`https://www.ijcai.org/Proceedings/1971`) indicate the breadth AI had achieved by 1971. Alan Turing, Herbert Simon, John McCarthy, and Marvin Minsky had previously recognized its deep relevance for philosophy, including philosophy of mind. Arguing for the heuristic inadequacy of pure logic-based AI, Sloman (1971) offered a new defence of Immanuel Kant's philosophy of mathematics, summarised in Sloman (1965), claiming that some kinds of mathematical knowledge are (a) *non-empirical*, (b) *synthetic/non-analytic* i.e. not based merely on logic and definitions and (c) include *necessary* (= non-contingent) truths.

An important potential (future!) use of AI is explaining why Kant's philosophy of mathematics was broadly correct, especially about discoveries concerning constructions and proofs in Euclidean geometry—contrary to popular opinion among philosophers and mathematicians who think Kant was refuted by Einstein's theory of General Relativity, and Eddington's observation of the 1919 solar eclipse, as argued in Hempel (1945).[4] A future AI system making mathematical discoveries with the features described by Kant, might replicate in a "baby robot" the ability of some baby humans to grow up to be mathematicians. This will require deep advances in biology, neuroscience, and philosophy, as well as AI. Sloman (1962) offered a purely philosophical defence of Kant that would be considerably strengthened by advances in AI replicating human and non-human spatial reasoning competences.

3. Challenging representational constraints in AI

Despite McCarthy's and Hayes's claims for adequacy of logic-based forms of representation for AI, it is arguable that if ancient mathematicians had been restricted to exploring what can be done using logic they would not have discovered the constructions and proofs in Euclidean geometry

2. A very brief, incomplete, introduction to the ideas can be found in `http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL\_COPIES/OWENS/LECT8/node2.html`
3. For more on the work of Max Clowes see the obituary notice and bibliography (Sloman, 1984 to 2018).
4. Also at `http://www.ditext.com/hempel/geo.html`

that are still in use world wide. Rather than logic being *heuristically* adequate, being restricted to using and thinking with logical forms of representation would have made ancient discoveries much harder than the use of diagrams and diagrammatic constructions (including *imagined* diagrams and constructions). Although some theorem provers can prove theorems in Euclidean geometry e.g. Gelernter (1964) and the far more sophisticated Chou, Gao, and Zhang (1994), they work only because their designers provided logicised versions of Euclid's axioms and postulates e.g. Hilbert (1899), which the original ancient geometers did not have and did not need: they used other, still unknown, mechanisms for studying spatial structures and processes. Sloman (1962) defended the validity of ancient diagrammatic forms of reasoning, without reference to AI. Future AI and neuroscience, explaining the roles of sub-neural chemistry in spatial reasoning in brains, may produce a much better defence of Kant.

Similar remarks can be made about mechanical engineers designing or debugging complex machines with 3D interacting parts, such as gears (including worm and pinion gears), pulleys, levers, cables, pistons, etc. Has any engineer tried designing a functioning crane or other complex piece of machinery, using only predicate calculus (plus modal logic if needed) to describe the structures, their relationships, their functions, and the processes that can occur during their operation? A computer might be programmed to do it using only logic and arithmetic, but it would not be an accurate model of human design processes, if it replaced all spatial reasoning by numerical and logical reasoning. Moreover, it is very unlikely that replacing all the spatial toys used by pre-verbal children and trying to teach them logic, and formal versions of Euclid's axioms instead, will increase their spatial understanding and future powers as scientists, engineers, architects, or carpenters. Neither would replacing their chemistry-based brains with statistics-based neural nets if that were possible.

Likewise, I suspect that replicating ancient mathematical discovery processes, and also everyday processes of spatial reasoning, cannot be done on digital computers, whether they use logical theorem provers or artificial neural nets, if brains make essential use of sub-neural chemical processes with a mixture of continuous and discrete changes.[5] In contrast, neural net models using statistical evidence to derive probabilities, cannot even *represent* impossibility or necessity, let alone find proofs of impossibility or necessity. Neither can neural nets in brains, for the same reason, which suggests that understanding ancient mathematical discovery processes will require an understanding of how brains use sub-neural chemical mechanisms, with a mixture of continuous and discrete processing, which I suspect motivated the research reported in Turing (1952), very different from Turing's earlier work on Turing machines Sloman (2002).

Some neuroscientists are now investigating sub-neural computations for other reasons, e.g. Trettenbrein (2016); Grant (2018). Perhaps 22nd Century (or later) AI system will use mechanisms that are now unimaginable: one of the themes of the Turing-inspired "Meta-morphogenesis" project.[6]

5. I have several partially analysed online examples, and would welcome help with making further progress, e.g. `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/deform-triangle.html`, `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/super-turing-geom.html`
6. `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`

4. Symbolic, logic-based AI

One of the less-visible, less-fashionable, major strands in current AI inspired by the early work of McCarthy and others is the use of logic, algebra and arithmetic for reasoning and discovery. There are powerful theorem provers used in practical applications such as proving termination of programs, or satisfaction of formal requirements (subject to adequate physical memory and time limits), e.g. `https://www.embedded.com/you-think-your-software-works-prove-it/`. Such definite conclusions cannot be reached by statistics-based learning systems or any mechanism whose results always have attached probabilities.

When we fully understand human spatial reasoning mechanisms and their roles in ancient mathematical discoveries, we may not be able to replicate them in current computer-based systems, in which case AI will have to be expanded to include the study of biologically evolved computational mechanisms, perhaps including sub-neural chemical computations, a possibility requiring further research. This would render out of date many 20th and 21st century specifications of what AI is.

Finally, this discussion presupposes notions of information and information processing. But I am not referring to Shannon information introduced in (1948), which is basically a *syntactic* property. Instead I have been using the much older *semantic* concept of information, used, for example, in Jane Austen's novels a century before Shannon: `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/austen-info.html`, a far more important concept for organisms or machines perceiving, interacting, and learning in a complex, richly structured, constantly evolving environment.

References

Abercrombie, M. 1960. *The Anatomy of Judgement*. New York: Basic Books.

Boden, M. A. 1977. *Artificial Intelligence and Natural Man*. Hassocks, Sussex: Harvester Press. Second edition 1986. MIT Press.

Boden, M. A. 2006. *Mind As Machine: A history of Cognitive Science (Vols 1–2)*. Oxford: Oxford University Press.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chou, S.-C., Gao, X.-S., and Zhang, J.-Z. 1994. *Machine Proofs In Geometry: Automated Production of Readable Proofs for Geometry Theorems*. Singapore: World Scientific.

Clowes, M. B. 1971. On seeing things. *Artificial Intelligence* 2(1):79–116.

Clowes, M. B. 1973. Man the creative machine: A perspective from Artificial Intelligence research. In Benthall, J., ed., *The Limits of Human Nature*. London: Allen Lane.

Gelernter, H. 1964. Realization of a geometry-theorem proving machine. In Feigenbaum, E. A. and Feldman, J., eds., *Computers &Amp; Thought*. New York: McGraw-Hill. 134–152. Re-published 1995 (ISBN 0-262-56092-5).

Gombrich, E. H. 1960. *Art and Illusion: A Study in the Psychology of Pictorial Representation*. New York: Pantheon.

Grant, S. G. N. 2018. Synapse molecular complexity and the plasticity behaviour problem. *Brain and Neuroscience Advances* 2:1–7.

Hempel, C. G. 1945. Geometry and Empirical Science. *American Mathematical Monthly* 52. Repr in Readings in Philosophical Analysis, ed. H. Feigl and W. Sellars, New York: Appleton-Century-Crofts, 1949.

Hilbert, D. 1899. *The Foundations of Geometry*. Salt Lake City: Project Gutenberg. Translated 1902 by E.J. Townsend, from 1899 German edition.

McCarthy, J. and Hayes, P. J. 1969. Some philosophical problems from the standpoint of AI. In Meltzer, B. and Michie, D., eds., *Machine Intelligence 4*. Edinburgh, Scotland: Edinburgh University Press. 463–502.

Miller, G. A., Galanter, E., and Pribram, K. H. 1960. *Plans and the Structure of Behaviour*. New York: Holt.

Minsky, M. L. 1963. Steps toward Artificial Intelligence. In Feigenbaum, E. and Feldman, J., eds., *Computers and Thought*. New York: McGraw-Hill. 406–450. (Originally in *Proceedings of the IRE* 1961).

Minsky, M. L. 2006. *The Emotion Machine*. New York: Pantheon.

Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423 and 623–656.

Sloman, A. and Croucher, M. 1981. Why robots will have emotions. In *Proc 7th Int. Joint Conference on AI*, 197–202. Vancouver: IJCAI.

Sloman, A. 1962. *Knowing and Understanding: Relations between meaning and truth, meaning and necessary truth, meaning and synthetic necessary truth (DPhil Thesis)*. Ph.D. Dissertation, Oxford University.

Sloman, A. 1965. 'Necessary', 'A Priori' and 'Analytic'. *Analysis* 26(1):12–16.

Sloman, A. 1971. Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, 209–226. London: William Kaufmann. Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971.

Sloman, A. 1984 to 2018. Experiencing Computation: A Tribute to Max Clowes. In Yazdani, M., ed., *New horizons in educational computing*. Chichester: Ellis Horwood Series In Artificial Intelligence. 207–219. (Online version with expanded obituary and biography.).

Sloman, A. 2002. The irrelevance of Turing machines to AI. In Scheutz, M., ed., *Computationalism: New Directions*. Cambridge, MA: MIT Press. 87–127. `http://www.cs.bham.ac.uk/research/cogaff/00-02.html\#77`.

Trettenbrein, P. C. 2016. The Demise of the Synapse As the Locus of Memory: A Looming Paradigm Shift? *Frontiers in Systems Neuroscience* 10(88).

Turing, A. M. 1952. The Chemical Basis Of Morphogenesis. *Phil. Trans. R. Soc. London B 237* 237:37–72.

Journal of Artificial General Intelligence 11(2) 97-100, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-11-11
Accepted 2020-01-16

Intelligence Is Not One Thing

Alan Winfield ALAN.WINFIELD@BRL.AC.UK
Bristol Robotics Laboratory
University of the West of England
Bristol, UK

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

1. Introduction

Defining Artificial Intelligence is indeed problematical. A problem that arguably has it roots in the invention of the term Artificial Intelligence (AI). Moor (2006), for instance, speculates on "whether the field would have been any different had it been called *computational intelligence* or any of a number of other possible labels." But surely the problem is not with defining artificial intelligence, but natural intelligence. McCarthy's own definition of AI seems to me perfectly satisfactory: "it is the science and engineering of making intelligent machines" (McCarthy, 2007). What eludes us is a good definition of intelligence.

2. What Intelligence is not

I would like to challenge two common assumptions regarding natural intelligence. The first is that intelligence is a singular property of animals and the second is that intelligence falls on a linear scale from zero to superintelligence. Let me counter these misconceptions with Boden's elegant assertion (Boden, 2010) that "intelligence is not one thing that animals have more or less of."[1] Noting that deciding what something is not can be just as important as deciding what it is, let us develop Boden's insight.

If intelligence is not one thing then what kinds of things is it? In (Winfield, 2017) I propose that there are four categorically different kinds of intelligence, which I label as (1) morphological intelligence, (2) swarm intelligence, (3) individual intelligence and (4) social intelligence. Let me summarise these as follows.

1. Morphological intelligence is the kind of intelligence that a physical body confers to its owner. The idea that a body has some intrinsic intelligence may seem odd, but is closely related to the notion of morphological computation; which has been defined as "a term which captures conceptually the observation that biological systems take advantage of their morphology to conduct computations needed for a successful interaction with their environments" (Hauser, 2013).

1. Boden was not the first to observe that intelligence is not one thing. McCarthy (2007) makes the same point and Brachman (2006) gives an excellent account of the multi-facetted nature of intelligence, noting also that "intelligence is not created by just mixing together the individual facets."

2. Swarm Intelligence describes the collective, self-organised behaviour we observe in animals that swarm, shoal, flock or herd, or—more dramatically—build complex nest structures such as ants, bees and termites do. Swarm intelligence is an emergent property of the collective that results from the local interactions of the individuals with each other and with their environment (Dorigo and Birattari, 2007).

3. Individual intelligence is defined as the ability to both respond (instinctively) to stimuli and, optionally, learn new—or adapt existing—behaviours through, typically, a process of trial and error. If learning is present the actual learning mechanism is not important, except that it is the individual that learns in its own lifetime, without the help of another individual.

4. Social intelligence is the kind of intelligence that allows animals or robots to learn from each other. This might be through imitation or instruction. In imitation a new behaviour is acquired by the social learner observing another's behaviour then transforming those observations into corresponding actions and responses.

In (Winfield, 2017) I assert that animals and robots do have more or less of each of these four kinds of intelligence, and suggest a way of graphically comparing the intelligence of different animals and robots on radar charts with four axes, one for each kind of intelligence (while also admitting that making quantitative estimates of each remains very difficult). This approach does nevertheless provide useful insights into the reasons for the chronic intelligence deficit of present day intelligent robots when compared with animals.

Let me now reflect on the importance of morphological intelligence. I regard embodiment to be important in any discussion on both natural and artificial intelligence, for two reasons: first the simple fact that without exception every example of natural intelligence we know of is embodied, and second, that embodiment shapes the way all animals, including us, think and behave (Pfiefer and Bongard, 2006). Even the smartest human cannot fly unaided like a bird: his body simply does not have the morphology or morphological intelligence. When we do (ingeniously) construct flying machines we are literally (to use Dawkins' metaphor) extending our phenotype (Dawkins, 1982); a minimal flying machine such as a microlight, for instance, equips a human with the artificial morphological intelligence to fly.

As a final reflection on embodiment, if we accept as the originators of AI did, that the true aim of AI is to simulate not intelligence in general but human intelligence in particular, then those AIs will need to experience the world in the same way as we humans, and that is only possible if they have bodies able to act and interact in the human world. As Dreyfus (1996) argues, "...one would need to have experience with our kind of body to make sense of our kind of world."

3. Intelligence and Adaptation are not the same thing

A straightforward and generally uncontentious definition of intelligent behaviour is *doing the right thing at the right time*. But its simplicity is deceptive, for in order for an agent to do the right thing it first needs to determine what the right thing is. This is not simply a matter of selecting an action from a pool of next possible actions (Seth and Bryson, 2013). In order to select the right action the agent needs to be able to perceive the context within which it finds itself with sufficient precision to be able to make a determination of which action is best. And if the situation is dynamically changing then that determination may well also require the agent to be able to internally model itself and the world in order to anticipate the future (Winfield and Hafner, 2018).

Action selection therefore requires an agent to be able to perceive context, quite possibly with limited or noisy sensor information, then determine which action is best, without equivocation, when no single action emerges as optimal (think of an animal having to choose whether to fight or flee when she and her antagonist seem well matched, or a chess player having to choose her next move against the clock). These are sophisticated capabilities that surely deserve to be labeled as intelligent. Yet they do not—at the moment they are executed—involve adaptation.

I contend that there is value in separating definitions of intelligence and adaptation, as follows.

- An agent can be described as intelligent if it is capable of determining the most appropriate course of action for the situation in which it finds itself, in the presence of both uncertainty and incomplete information, and then executing that action.

- Adaptation is the ability of an intelligent agent to acquire new strategies for action selection and/or new actions to select.

Separating intelligence and adaptation[2] in this way has the additional merit of highlighting the temporal and energetic differences between intelligence and adaptation: intelligent behaviour must be timely and energy efficient, whereas adaptation is a slower and more energetically costly process.

In conclusion I believe that one of the reasons we still have such difficulty defining intelligence is that, as Moor (2006) writes in his report of the AI@50 conference held to mark the 50 year anniversary of the foundational Dartmouth conference, "...there still is no general theory of intelligence or learning that unites the discipline."

References

Boden, M. 2010. Personal communication.

Brachman, R. J. 2006. (AA)AI—more than the sum of its parts, 2005 AAAI Presidential Address. *AI Magazine* 27(4):19–34.

Dawkins, R. 1982. *The Extended Phenotype*. Oxford University Press.

Dorigo, M. and Birattari, M. 2007. Swarm Intelligence. *Scholarpedia* 2(9):1462.

Dreyfus, H. L. 1996. Response to my critics. *Artificial Intelligence* 80(171–191).

Hauser, H. 2013. Morphological Computation and Soft Robotics, ShanghAI lectures. `https://shanghai-lectures.github.io/archives/sites/default/files/guestlectures_slides/FINAL_ShanghAI_lecture_2013.pdf`.

McCarthy, J. 2007. What Is Artificial Intelligence? `http://www-formal.stanford.edu/jmc/whatisai/`.

Moor, J. 2006. The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine* 27(4):87–91.

2. Of the four kinds of intelligence outlined in Section 2, on reflection two: individual and social intelligence are actually kinds of adaptation.

Pfiefer, R. and Bongard, J. 2006. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press.

Seth, A. K. and Bryson, J. J. 2013. Natural Action Selection, Modeling. In Pashler, H., ed., *Encyclopedia of the Mind*. Sage. 557–559.

Winfield, A. F. T. and Hafner, V. V. 2018. Anticipation in Robotics. In Poli, R., ed., *Handbook of Anticipation*. Springer.

Winfield, A. F. T. 2017. How Intelligent is your Intelligent Robot? arXiv:1712.08878 [cs.RO]. https://arxiv.org/abs/1712.08878.