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.