The Future of AI Research: Ten Defeasible 'Axioms of Intelligence'

Kristinn R. Thórisson THORISSON@{RU.IS, IIIM.IS}
Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland
Icelandic Institute for Intelligent Machines, Reykjavik, Iceland

Henry Minsky HQM@LEELA.AI
Leela AI, Cambridge, Massachusetts, USA

Editor: Kristinn R. Thórisson

Abstract

What sets artificial intelligence (AI) apart from other fields of science and technology is ***not*** what it has achieved so far, but rather what it set out to do from the very beginning, namely, *to create autonomous self-contained systems that can rival human cognition*—machines with 'human-level general intelligence.' To achieve this aim calls for a new *kind* of system that, among other things, unifies – in a single architecture – the ability to represent causal relations, create and manage knowledge incrementally and autonomously, and generate its own meaning through empirical reasoning and control. We maintain that building such systems requires a shared methodological foundation, and calls for a stronger theoretical basis than simply the one inherited directly from computer science. This, in turn, calls for a greater emphasis on the theory of intelligence and methodological approaches for building such systems. We argue that necessary (but not necessarily sufficient) components for general intelligence must include the unification of causal relations, reasoning, and cognitive development. A constructivist stance, in our view, can serve as a good starting point for this purpose.

Keywords: Artificial intelligence, AI Methodology, General Machine Intelligence, Artificial general intelligence, Human-level intelligence, Constructivist AI, Constructivist Psychology, Cumulative Learning, Machine Learning, Knowledge Representation, Autonomy, Generality, Autonomous Generality

1. The Aim of AI Research

Since its inception as a separate field of science and technology, the field of artificial intelligence (AI) always had an ultimate basic research goal (cf. Simon (1995); McCarthy (1983)): To explain the phenomenon of intelligence in sufficient detail to allow the creation of machines with that very property. Frequently discussing what is often referred to as 'human-level intelligence,' the founding fathers of AI used this term in the 1950s and '60s in precisely the same way as it is used today to describe general-purpose cognitive skills—autonomous agents capable of learning a variety of things in a variety of environments, much like a good human employee.

While the public discourse about AI has always included a heavy dose of science fiction in its narrative (the 'real AI' that seems perpetually just-around-the-corner), historically its primary focus has been squarely on near-term practical applications. This seems to have been the case from the very beginning (except perhaps at the earliest stages); in 1983, John McCarthy, then president of the American Association for Artificial Intelligence,[1] wrote in the President's Quarterly Message:

> AI NEEDS MORE EMPHASIS ON BASIC RESEARCH
> *Too few people are doing basic research in AI relative to the number working on applications. The ratio of basic/applied is less in AI than in the older sciences and than in computer science generally. This is unfortunate, because reaching human level artificial intelligence will require fundamental conceptual advances.*
> — J. McCarthy (1983, p.5)

The difference between technological application-guided research and scientific (question-guided) research boils down to what kind of explanation is being sought by the effort: Basic science seeks the fundamental reasons and principles for why things are the way they are, while technological development and application primarily seeks to explicate which solutions can be applied to which ends. While the latter may bring forth many innovations, it very rarely (if ever) invents new paradigms or uncovers fundamental underlying principles—this only happens through a concerted effort focusing exclusively on open questions. This is why only science-focused research can move the field of AI consistently and systematically forward.

To be sure, artificial intelligence research has contributed to both scientific and technological progress. Rather than contributing to the development of *theories of intelligence*, however, in active dialog with cognitive science, neuroscience, or psychology, the field's key contributions have been largely limited to computing (i.e. mathematics) and computation theory proper (cf. Waltz (1999)), with highlights including e.g. informed search, game play, speech recognition, computer vision, and more. It does not take an enormous stretch of the imagination to see that most, if not all, of the touted advances that work in AI has produced could have sprung forth without a special focus on AI per se; in an alternate universe, time-sharing, object-oriented programming, A* search, and other products typically classified as the fruits of AI-focused R&D could very well have been produced by straight-shooting computer science researchers only mildly inspired by natural intelligence.

Being so intertwined with and seemingly inseparable from computer science, a question inevitably arises: Why do these contributions deserve a special label, 'AI'—why aren't simultaneous localization and mapping methods, computer vision approaches, robot control architectures, and speech recognition techniques, simply considered part of the computer science curriculum, plain and square? This question has been so persistent and nagging as to eventually having been given its own name – "the AI effect" – which states that as soon as a computational solution has been found for a particular sub-problem or task in AI, and which before was considered to require intelligence, it is re-classified a 'computational problem' with a known 'computational solution,' in no need for 'intelligence.' The prototypical and best-known example being chess. So, what – if anything – is special about AI?

1. Founded in 1979, the name of AAAI was changed in 2007 to 'Association for the Advancement of Artificial Intelligence.'

2. What is Special About Artificial Intelligence

There are undoubtedly many ways to define intelligence in ways that help distinguish it from other fields of research, including computer science, but fewer definitions may exist that lead to a clear, concise, and unequivocal separation of AI from the full set of psychology, biology, neurology and computer science. Here we will highlight one that meets that criterion: *Cognitive autonomy*.

Autonomous general learning – i.e. general self-supervised learning – involves the *autonomous creation* of knowledge structures about phenomena *unfamiliar* to a learner (cf. Swan et al. (2022); Wang (2020); Sheikhlar et al. (2020); Thórisson (2020)). This is a process involving actions, measurements, and systematic construction of structured and compositional information; the knowledge thus generated by a learner – without outside teaching assistance of any kind – is produced through an interaction between the learner's knowledge acquisition mechanisms and its environment, resulting in compositional information structures that builds up explainable perception, explainable action, and explainable learning.

A theory of cognitive autonomy – the ability of a system to learn to operate in an environment starting with only general operating principles and knowledge at 'birth' – should be a fundamental focus in AI (cf. Swan et al. (2022)). In 1999 the President of AAAI David Waltz asked: *"The study of intelligence is an ill-posed problem: What is intelligence? A better question is What is intelligence for?"* (Waltz, 1999, p.20). Intelligence is not just one approach (of many) to achieve the *immediate application* of actionable information to (complex) problems,[2] it is in fact *necessary* for *the production of new solutions to new problems and situations* in light of active goals (cf. Wang (2020)). This has sometimes been called 'common sense' (cf. Panton et al. (2022)) but we prefer to call it simply 'understanding' (Thórisson and Kremelberg (2018); Thórisson et al. (2016)).[3]

Understanding a task cannot simply mean the ability to achieve it, because then any machine that performs according to its specifications would 'understand' (does a thermostat really *understand* room temperature? We think not). No, to really understand something means not only the ability to achieve goals with respect to it (Thórisson et al. (2016)) but also an ability to explain any relevant aspects of that something (cf. Thórisson (2021)).

> The task of AI research is to enable the creation of systems that can do this *of their own accord*: Formulate *new solutions* to problems, new and old alike, and *explain the principles* behind those solutions.

Or, more precisely, the top-level task of *scientific* progress in AI involves analyzing and explicating the principles by which a system *becomes* intelligent; the top-level task of AI

2. A common definition of AI used by its researchers is "achieving complex goals in complex environments" (cf. Legg and Hutter (2020)), yet it is not very clear what is meant by "complex" and "problems."
3. This does not mean that 'understanding' and 'common sense' should be considered synonyms, quite the contrary (Thórisson and Kremelberg (2018)); the latter concept brings with it a lot of anthropomorphic implications (that has diffused the discussion of these phenomena and probably slowed down progress in AI for decades). 'Understanding' seems, to us, not simply a less burdened term but also more obviously related to general intelligence than "common sense," which neither seems sensical nor very common.

engineering is to apply those principles in the creation of useful intelligent systems.[4] What AI researchers should focus on, over and above anyone who is not interested in AI specifically, is the creation of machines that can operate on their own accord, autonomously: learn by themselves, seek explanations by themselves, and thus deepen not only their skill set but also their own understanding.

> What is special about artificial intelligence is thus, ***not*** what it has achieved so far, but rather *what it set out to do*, namely, *to create autonomous self-contained systems* that can rival human cognition—*'human-level intelligence.'*

At its core, the field of AI strives towards a deeper understanding of the phenomenon of intelligence. This cannot be done by slicing and dicing its subject matter into bits and pieces and working on each one completely independently of the others; there must be a concerted effort to include all necessary and sufficient ingredients to create such machines. Although the ability to classify is necessary for intelligence, to move AI beyond its current state and enable machines that can produce new solutions for new problems – and do so increasingly autonomously – we must also include *control* in our set of necessary topics. This requires knowledge of cause and effect (cf. Thórisson and Talbot (2018)), because effective control in complex task environments cannot be achieved based on solely correlational information. This necessitates appropriate machinery to enable an AI to represent cause and effect, and use this a key component in its reasoning processes, including hypothesizing about the causal relationships of novel phenomena.

Unifying the necessary and sufficient set of system properties, to create machines like these, is virtually impossible without a proper theoretical foundation. No amount of wishful thinking can bring isolated sub-systems together, built on incompatible theoretical assumptions, following differing methodological approaches. Systems built in a piece-wise, feature-isolation manner inevitably end up with severe shortcomings in their cognitive functionality.

This was McCarthy's (1983) call to arms, and what Newell (1994) continued a decade later to call for. To move AI forward we need more basic research—a greater emphasis on theory, cognitive autonomy, and proper methodologies. Without a renewed focus on these, the field will continue to move along at a snail's pace, continuing to relegate real AI to science fiction.

3. Ten Defeasible "Axioms of Intelligence"

We have identified ten important tenets that summarize our view on the subject, to more generally help set an agenda for AI research in the coming decades. These tenets can be thought of as working hypotheses about the phenomenon of intelligence, formulated in a way intended to help those who are interested in working towards general machine intelligence and break the current stand-still when it comes to common theoretical and methodological foundations for the field. These are not the "ten commandments of AI,"—they should be debated, scrutinized, improved, and eventually replaced, once something better comes along.

4. A frequent confusion between the scientific pursuit of artificial intelligence and the industrial application of intelligent systems has lead to a muddled and unfocused discourse between the media and the public; see for instance Ray (2022) for a clear example of this.

§ H0 *Intelligence is a systemic phenomenon, requiring the unification and coordination of many known and undiscovered information processing principles.*

Creating intelligent machines is a scientific and engineering challenge, requiring both theoretical and technical advances in autonomous systems design. There exists no single principle or "golden algorithm" that will "solve intelligence" and make it possible to reach "human-level" intelligence or general machine intelligence in one fell swoop (contrary to what some research has hypothesized, cf. Silver et al. (2021)). "Solving intelligence" will require discovering and unifying many principles and ideas that must be implemented in a single system.

§ H1 *To achieve autonomy, and be capable of general learning, an agent in the physical world must be able to learn incrementally and modify its existing knowledge in light of new information.*

If we want our future AIs to operate effectively and efficiently, independently, in the physical world, learning must be incremental. For an intelligent agent acting in a complex ever-changing environment, no hypothetical or realistic scenario exists under which every conceivable piece of potentially relevant information is available to the agent up front—whether before entering the world or any time thereafter. This extends to *any and all information:* perceptions, actions, predictions, implications, plans, and so on. Learning has traditionally been divided up into a variety of smaller bits and pieces; the concept of ongoing (always-on) learning has been called many names, including 'lifelong', 'perpetual', 'never-ending', 'incremental', 'online', and 'continual' learning (cf. Mitchell et al. (2018); Zhan and Taylor (2015); Zhang (2018); Fontenla-Romero et al. (2013); Silver et al. (2013)), based on definitions of terms which have partial overlap (at best) and are invariably are based on ontologically incompatible principles. Old-new unification is in our view a key concept for cumulative learning, but has seldom been put in the foreground of AI research, with very few exceptions (cf. Swan et al. (2022); Nivel et al. (2013)).

We see cumulative learning as a *unified mechanism* that combines online/always-on, incremental, reasoning-based acquisition of increasingly useful information about how things work.[5] The requirement of unification is important, because from this unification comes its power and potential for enabling general intelligence. Cumulative learning is not yet a major focus in AI research.

§ H2 *To enable incremental buildup and modification of knowledge, cumulative (i.e. incremental, continual) learning requires compositional knowledge representation. When facing novel phenomena, such knowledge creation must be based on informed, contextual, focused, and defeasible* (Pollock (2010)) *hypothesis generation.*

5. Our definition of 'cumulative learning' follows Thórisson et al. (2019; 2018). The term has appeared elsewhere (cf. Chen and Liu (2016); Fei et al. (2016); Baldassare et al. (2009)) with partial overlap in definition.

The kind of compositional knowledge representation we envision would mirror the compositional nature of the physical world, making it relatively straight forward, conceptually at least, to see how a modeling process unfolds during learning, where logic is used to unify new information with old, and less useful models are improved when new evidence for their improvement becomes available. The appropriate representation would support easy comparisons of wholes and parts, reducing storage requirements for any self-similar environment, lending support to features of a human mind like analogy making.

Swan et al. (2022) and Thórisson et al. (2020; 2019; 2018) describe frameworks for autonomous cumulative modeling that allows an artificial agent to generate causal networks automatically, through experience, that can be evaluated along the above dimensions, and subsequently use them in its further learning. The approach has been tested on complex tasks involving human-robot interaction (Thórisson et al. (2014)) and as of yet, is the only known approach demonstrating autonomous general learning along these lines.

The handling of novelty requires hypothesis generation, whether explicit or implicit.[6] Informed hypothesis generation is a necessary (but not sufficient) mechanism for dealing with the unexpected (the most expensive method for handling it is random search—which doesn't scale for a general learner in a complex world). How this informed hypothesis generation can be implemented is an important topic for future AI research.

> **§ H3** *To keep track of arguments for and against knowledge hypotheses, a capacity for (self-)explanation is necessary. This explanation capacity is based on reasoning processes, including deduction, abduction, induction, and analogy.*

Given a particular perceived circumstance (i.e. measurement), desired target state, fact, situation, or process, an explanation for its validity (e.g. in light of active goals) may be generated through a process of logical argument, whereby seemingly relevant but competing models of causal (and other) relations are compared and contrasted, in an effort to identify the proper context with relevant given assumptions, that – if absent – would lead (or would have lead) to a different outcome. Keeping track of the history of such internal arguments (and recursively the argumentation itself) in which participating mental processes may be included, requires general methods for comparison, contrasting statements, and evaluation. This process is generally called 'explanation' (cf. Thórisson (2021)). It is also one of the main reasons why we hypothesize that reflection is a key component of any general intelligence (see **H5** below).

> **§ H4** *To achieve existential autonomy in a world with limited resources (time, knowledge, etc.), informed (self-)control of available resources is necessary, whether learned or pre-programmed.*

In world with a external clock, deadlines are common. To achieve complex tasks under these constraints calls for planning. Creating plans takes time, and since this step must

6. Explicit hypothesis generation would require reification of hypotheses and the processes used to compare them, allowing an agent to reflect on the hypothesis generation and evaluation using its learning mechanisms; an implicit approach would thus be simpler to implement, as it does not give the learning system explicit access to the generation and argumentation processes; see **H3**.

be done before the resulting plan is executed, it must either have a fixed predictable time, or a way to be characterized such that its percentage of the whole time available can be reliably predicted. In a system whose knowledge and cognitive machinery is developing and changing over time, this can only be done by modeling not only tasks and environments and learning from these but also modeling the cognitive processes at hand (Nivel and Thórisson (2013)). Modeling of what is *not* known at the time of planning is another important requirement. This, then, would provide the minimum information for sufficient (self-)control of planning and plan execution processes.

Another process worth mentioning in this context is memory access—bringing useful knowledge to a task at any point in time. For a narrow AI system this is typically not an issue, as such systems are assumed to be situated in their target role at all times; a general learner with a diverse set of top-level and sub-goals, however, could be facing a wide variety of situations, sub-tasks, unexpected events, etc. Accessing the most useful knowledge in a timely manner is absolutely key for such learners. This is intricately linked to the point about informed hypothesis generation mentioned in **H2**.

§ **H5** *If one or more of the above hypotheses* **H2** *to* **H4** *are correct, transparent operational semantics are necessary for general learning, as well as for achieving cognitive growth (see* **H7** *below). Achieving true autonomous artificial intelligence (including "human-level," "general machine intelligence," or any other reasonable conceptualization of a* truly *intelligent machine), is therefore highly likely to require reflection. Autonomous general learners must therefore be capable of reflection.*

The requirement for reflection comes from two primary sources. Firstly, because various *particular* thought processes at a particular time and place (such as your first thought when I tell you that 'summer is coming') and *categories* of thought processes (e.g. the class of predicting other people's utterances, or the class of methods for recalling what you yourself did yesterday), are regularly reified when we think about our actions and learn about the world (these made-up examples of thought process categories are a proof in point). Operations on reified thought serve an important role in our everyday planning and in growing our knowledge. Secondly, if the machinery of thought is going to develop – i.e. change systematically, based in part on experience, what we call 'cognitive growth' – the processes targeted for change must be measured and/or compared, calling for operations that require reification of (some of) their parts. On both counts, some kind of reflection is called for.

§ **H6** *Knowledge abstraction is a key feature of general autonomous cognition—without it, handling complex information at multiple levels of detail becomes intractable.*

Rather than programming AI systems with heuristics, we need systems that can come up with new heuristics on their own, when producing new solutions to new problems. People do this all the time, on demand (e.g. 'there is too much traffic downtown at this hour to take that route;' 'every time I use this browser my online order fails'), and after all, the concept of heuristics is itself a creation of our own minds. Our AI systems should also be

able to come up with concepts such as 'infinitely small point' and 'perfect circle.' This capacity cannot be programmed in the seed (i.e. provided a-priori; see **H7**)—to serve the same role as abstraction does in human thought, it must spring forth from first principles of the cognitive apparatus itself.

> **§ H7** *Existential autonomic learning means freedom from a teacher,*[7] *which means that the learner must be provided with a program (a 'seed'), up-front (at 'birth'), that contains actionable principles for bootstrapping its learning, and the development and growth of the cognitive apparatus over time. The more general this knowledge, the more flexible can the learning get, albeit at the cost of taking more time.*

Cognitive growth is the dynamically-steered development of the learning apparatus itself. Whether measured in hours, days, years, or decades, to be capable of such growth, a system must be able to program itself (cf. Nivel and Thórisson (2013)). Seed-programmed learning and cognitive growth have not been studied extensively in AI (Thórisson (2020)). Progress in research on autonomy and general learning depends on discovering principles for knowledge bootstrapping and growth control: How top-level goals (drives) relate to early learning, and how these develop as knowledge grows. Can the case of a 'newborn' be seen as a special case of *learning anything in light of novelty*?

> **§ H8** *A holistic stance and approach to the phenomenon of intelligence – that is, a research program that does not dismember the topic under study, transforming it in the process to a mere facsimile of itself – is by far more likely than other methodologies to deepen our understanding of intelligence and enable the creation of truly intelligent machines.*

Any complex system made up of many parts at several levels of detail unavoidably contains multiple non-linear relations between its parts. In the absence of a reasonably accurate blueprint, which is certainly the case for intelligence, they must be studied as a whole, lest we risk decoupling important relations and changing the system into something other than what we really intended to study.

> **§ H9** *A constructivist view, with an emphasis on self-guidance and self-originated meaning, is a useful methodological stance (and one of possibly only very few) to help focus on the issue of holistic systems and unify all of the above principles in a single system.*

The basic principle from which constructivist thinking sprung is very compatible with the need for a stronger focus on autonomy in AI research. Whether a constructivist stance is taken simply as inspiration, or more concretely as prescriptive, we consider a good theoretical and methodological foundation of utmost importance in AI research. To be sure,

7. We define existential autonomy as the ability of a system to act without dependence on explicit outside assistance, whether from a human teacher or, in the case of general machine intelligence, the system designer (Thórisson (2020)).

constructivism is not the only theoretically motivated approach aimed at allowing us to break out of the allonomic methodological mold (Thórisson (2012a)), but in our view few others seem as useful or obvious.

While the above ten tenets go a long way in communicating what we see as important points for the future of AI research, we want to mention two key foundations that we consider useful going forward. The first is control theory, the second is constructivist psychology. But first we will spend a few thousand keystrokes on the topic of theory.

4. The Need For Explicit Theoretical Foundations

The field of AI, in our view, should have the aim of producing solid engineering principles for creating *any kind of intelligent machine*, founded on proper *scientific theoretical foundations of thought and intelligence*, including how general learning, reasoning, and full cognitive autonomy can be achieved in unified ways in a single system. Since few (if any) other fields of research seek the creation of machines with such capabilities, the field of AI must make its explicit subject matter the *theory of intelligence*.

Perhaps because the majority of research to date has prioritized application over deep theory, the nature of intelligence has yet to be properly revealed in an encompassing scientific account. Few attempts at producing overarching and unifying theories of intelligence have come forth since the field's inception in 1956 (cf. Moor (2006)). The best-known exceptions (cf. Minsky (1988); Newell (1994); Anderson et al. (2004); Wang (2006)) show little overlap or synergy, as they each rest on philosophical and methodological grounds that remain to be unified through foundational laws or common principles. By 'explicit theoretical foundations' we mean targeted research on unification, in line what Newell (1994) and others have called for in AI, as perhaps best exemplified by the field of physics over the past 200 years, which saw the expansion of the idea of the atom into a rigorous system explaining forces, energy, friction, etc.[8]

Does all this mean we should start writing more philosophical papers? Papers describing implemented systems can certainly, at the very least, be provided with a more thorough theoretical context in general. Should we turn to speculation over implementation? Absolutely not! Progress of an empirical research field is dependent on both theory and experimentation. Achieving this clearly requires more than simply speculating—working, implemented systems must be built, accompanied by the theoretical foundations that enable and explain them. Any methodology we follow must include both theory and practice, with techniques for each one iteratively informing the other.

Commonalities in methodological approaches and tools will be difficult, if not impossible, to achieve and coordinate without a common theoretical foundation. This can only be built through a shared vocabulary and working definitions. To date, the field of AI has proceeded mainly by addressing cognitive features observed in humans and other animals as separate and unconnected; examples include reasoning, learning, prediction (anticipation), planning,

8. Cf. https://en.wikipedia.org/wiki/Atomic_theory – accessed Nov. 21st, 2022. Note that our call to physics does ***not*** translate to letting mathematics dictate our theories (we cannot let the tools lead the way!). Only a solid philosophical foundation can provide the stability needed for formalization to prosper; we must avoid the pitfalls of premature formalization (Thórisson (2013)).

and perception, while largely ignoring others, e.g. resource control (attention), imagination, understanding ("sense-making"), reflection, and cognitive growth. These must ultimately all be brought together in a single theory.

The aim of AI research should neither be exclusively (or even mostly) speculation, nor limited to building simplified simulacra of isolated cognitive functions. Any theoretical construct aimed at advancing our understanding of how to implement unified cognitive functions and control – captured in a single system – should ultimately be judged on its potential to guide implementation: systems implemented according to the theory should allow us to conclude, through appropriate means, that it can lead to reasonably complete AI architectures that explain, in a unified way, a large portion of observed characteristics of general intelligence. 'Appropriate means' here involves of course experimental evaluation in target environments, especially that these systems can scale up to complex, real-world situations. With respect to autonomy, scaling up includes decreasing a system's need for calling home—for e.g. being reprogrammed or re-engineered in part by its developers; with respect to learning, scaling up means the ability of the systems to continue learning, resiliently, a variety of tasks in a variety of environments in light of a variety of obstacles, and efficiently and effectively handle novelty. It probably also involves the ability of these systems to explain and hypothesize about the world, themselves, and interactions between these.

To achieve the kind of autonomy that we envision, a learning system must be able to *create its own meaning* from *its own experiences* in the world. This viewpoint is a central theme of the constructivist learning school of thought in psychology (Piaget (2013)), and should therefore be of great interest in general intelligence: it forces us to think about how a system can manage its own cognition, including perceiving, making sense of, learning, and planning. This cannot happen, of course, without the system knowing (or learning) what to pay attention to—and when, how things hang together, and – perhaps most importantly – how to handle novelty.

5. The Role of Control

To get anything done requires some form of *control*, a principle and concept that has not exactly been front and center throughout AI's research history. To address control systematically includes characterization of the kinds of worlds, environments and tasks that we want the AI to handle—this defines the outer bounds of a system's operation, both physical and cognitive. Any moderately complex task-environment that includes several levels of spatio-temporal detail will provide enormous combinatorics—in this respect, the physical world can be assumed to present infinite compositionality. Most of these compositions cannot be documented (by *any* means) beforehand, and much of it cannot even be *foreseen* a priori. An autonomous learner must be able to create knowledge about such things on its own, and reason about them.

It's not just the manipulation of the immediate surroundings that an AI must have a handle on, the acquisition of knowledge itself also requires some form of process control, including sensory uptake, filtering, and processing of sense data; the unification of new knowledge with old knowledge requires control (a problem that gest exponentially larger with the amount and diversity of knowledge); the application of actionable knowledge re-

quires control; the management of resources in light of limited time, knowledge and energy requires control; the use of acquired knowledge for making plans requires control; learning to classify requires control; the creation of novel analogies requires control; the generation of hypotheses about new phenomena requires control: in short, everything that a generally intelligent agent may want to do in the physical world requires some form of control. To make progress towards general autonomous learning systems, these forms of control must be capable of (meta-)coordination.

We need shared methodologies – unifying principles and techniques used by communities of researchers – that allow experimentation in implementing autonomous learning and cognitive control across a reasonably wide range of ideas. A constructivist stance presents a useful, and to some extent unifying, starting point when posing questions related to representation, e.g. what kinds representations can be used for autonomnous learning about new environmental phenomena, by emphasizing the need for autonomic processes (in essence, cognitive metaprocesses). It also bears on the form and function representations need have to be amenable to online realtime automatic modification and manipulation in support of cognitive growth—autonomous self-programming.

A key functionality that such representations must capture is the ability to model cause and effect. Acting efficiently and effectively in a world where data is overabundant and processing power is scarce (what Wang (2009) calls an 'assumption of insufficient knowledge and resources' – AIKR) requires knowledge of causal relations. Here we need to beef up current efforts to cover the "last mile" from human-level intelligence (e.g. Pearl's (2012) do-calculus) down to the autonomic (informed, self-managed) creation of hypotheses about causal relations, and verification of these from experience (cf. Sheikhlar et al. (2021); Thórisson and Talbot (2018)).

Compared to their natural counterparts – besides a long list of missing features – the greatest shortcoming of all techniques, technologies, and systems produced by AI research so far is this: their inability to achieve independence from their creators. This is due to their lack of methods for *autonomous control*. The vast majority of programming languages developed to date, and paradigms created for software development in general, require human-level intelligence from the outset – the very phenomenon we are striving to figure out how to implement. Self-guided learning and cognitive growth – autonomic seed-programmed learning – requires transparent operational semantics to enable self-programming; how to achieve this is yet another question that begs to be answered. Questions regarding theoretical scalability issues loom large, and for this we need appropriate methodologies.

6. Which Methodology?

The key overarching methodology employed at present in AI research is inherited wholesale from computer science and software engineering, with the same exact programming methods and "algorithmic thinking" leading the way (Thórisson (2017)). Like in any standard software development, AI researchers build their knowledge-based systems by hand, like construction workers laying down the bricks of a house. This constructionist (not to be confused with constructivist) approach leads system development along the path familiar to every software developer: the task to be performed, and its execution environment, is defined and dissected by the human designers, solutions to it are proposed (by humans),

algorithms for dealing with them subsequently developed (by humans), and the resulting system behavior evaluated (by humans), when *these* are the very tasks that any generally intelligent system should be expected to do.

Virtually all available methodologies for AI (and they aren't very numerous) are of the constructionist kind. What is lost in this approach, among other things, are principles for self-adaptation, self-control, and self-generated meaning—in short, existential autonomy. Can something be done to beef up our methodological stance on these matters?

Methodological considerations are co-dependent on theoretical ones: If stars are 'holes in the heavens' we might find no reason to develop better telescopes; if creatures too small to see with our own eyes are 'a mere fantasy,' we see no need to invent a microscope. With a relentlessly myopic focus on practical ideas, many researchers in AI may not see much opportunity – or reason – to discuss methodology at all. The choice of an appropriate methodology is, however, a primary determinant of scientific progress, so there is a lot to be gained by starting off on the right foot.

While we wait for an accepted holistic theory of intelligence, an increased focus on methodological foundations could help bootstrap a new phase of research on general (machine) intelligence. The indications that initially helped researchers settle on a theory of microscopic creatures were much more coarse-grained than the phenomenon itself: statistics describing increased numbers of infections when certain procedures for cleanliness were ignored. In other words, the methodology came first, based on a rudimentary analysis of the phenomenon at hand. Similarly, we may be able to make some sensible methodological choices in AI by analyzing the key features of our subject matter. A good start is to list the requirements for the phenomenon we are interested (cf. Sloman (2000); Laird and Wray III (2010))—the features of the phenomenon that we consider necessary and sufficient.

Whatever our methodology involves, it should include approaches and techniques for representing causal relations, as already mentioned, explicit manipulation of subsets of knowledge, and its compositionality. Some might be tempted to call this a 'symbolic' approach (Nilsson (2007)), while we see it as both less than that and more than that: less, in that the term 'symbolic' tends to come with all kinds of (assumed) inherited features and assumptions from good-old-fashioned AI (Haugeland (1985)), many of which are outdated and irrelevant for systems that learn on their own accord, and more in that the knowledge of general learners must go beyond present approaches by allowing autonomic dissection, analysis, and manipulation of self-acquired knowledge, along the lines of what Abel (2008) calls 'the cybernetic cut.'

To handle heterogeneous task-environments in an informed and systematic way requires reasoning. A persistent focus on statistics-based knowledge representation throughout the history of AI (cf. Rosenblatt (1958); Sanger (1994)) has relegated reasoning and symbolic representation to the fringes of the field and making its comeback all but an afterthought (cf. Garcez et al. (2015)).[9] An agent with the proper hierarchical and causal knowledge data structures can construct useful new knowledge by trying to answer the right questions, if it

9. We see this as a clear example of letting the tools lead the way, at the cost of pushing fundamental questions aside. No good theoretical arguments can be found for why our approaches to general intelligence should require artificial neural nets that are so different form natural neural nets (Schaeffer et al. (2022)) and when there is no solution in sight for how they could break out of the human-centric methodological (constructionist) approach and into an autonomic one (Thórisson (2012b)).

can effectively join together existing high-quality islands of knowledge it already has constructed internally (cf. Nivel and Thórisson (2013); Mueller and Minsky (2015); Thórisson (2020)), using inference and mental simulation ('sub-activation' in Drescher's (1989) terminology). We consider this one of the highest levels of reasoning processing worth aiming for in a generally intelligent system.

In line with the conceptual foundations of constructivist psychology, we see constructivist AI (Drescher (1989); Thórisson (2012a)) to be a theoretical/methodological stance that seeks to create increasingly autonomous systems that can learn on their own from experience, generating their own knowledge without outside help, and modify their own internal structures to improve their own operational characteristics, based on experience. Thórisson (2012a) proposes principles for extending a constructivist viewpoint to serve a broader basis for creating appropriate conceptual tools to address the many challenges standing in the way of a unified foundational methodology for studying and creating general machine intelligence.

7. Conclusions

As we have argued elsewhere (cf. Thórisson and Helgason (2012)), future general machine intelligence (GMI), to deserve the label, must be capable of learning a wide variety of tasks and adapt to a wide range of conditions, none of which can be known at design time. This requires some minimum level of existential autonomy – the ability of a system to act without needing explicit outside assistance, whether from a human teacher or, in the case of GMI, the system designer. Existential autonomy calls for unplanned and unforeseen changes to a system's own cognitive structures; designing systems capable of this means we need appropriate methodologies for imparting this second-order control. To be capable of cognitive growth – whether measured in minutes, days, years, or decades – involves yet another level of control. Both categories of machines must ultimately be able to program themselves. Provided with (minimal) bootstrap knowledge, general machine intelligence involves systems that operate in in a way that, while initially limited in their knowledge and abilities, are capable of facing novel situations in their environment – simple at first – and grow their intelligence as experience accumulates. Given scalable principles, a system will continuously grow to ever-increasing levels of cognitive sophistication. The creation of such systems requires both theoretical and methodological advances beyond the present-day; the sooner we address these issues head-on, the sooner will we see machines with real intelligence.

Key ingredients that we see as necessary (but not sufficient) that we have emphasized in this paper, are the ability to handle novelty, the ability to manage experience autonomously, and the ability to represent causal relations. In our view, the last one has still some way to go, while a constructivist approach to AI provides a useful starting point for addressing the first two.

Acknowledgments

We would like to thank xx for useful comments on this paper. This work was supported in part by the Department of Computer Science at Reykjavik University, the Icelandic Institute for Intelligent Machines, a grant from Cisco Systems Inc., and Leela AI.

References

David L. Abel. The 'cybernetic cut': Progressing from description to prescription in systems theory. *The Open Cybernetics and Systemics Journal*, 2:252–262, 2008.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. In *Psychological Review*, volume 111, pages 1036–1060, 2004.

Gianluca Baldassare, Marco Mirolli, Francesco Mannella, Daniele Caligiore, Elisabetta Visalberghi, Francesco Natale, Valentina Truppa, Gloria Sabbatini, Eugenio Guglielmelli, Flavio Keller, and others. The IM-CLeVeR project: Intrinsically motivated cumulative learning versatile robots. In *9th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 189–190, 2009.

Zhiyuan Chen and Bing Liu. *Lifelong Machine Learning*. Morgan & Claypool Publishers, 2016. ISBN 1627055010, 9781627055017.

Gary L. Drescher. *Made-up minds: a constructivist approach to artificial intelligence*. PhD thesis, Massachusetts Institute of Technology, 1989.

Geli Fei, Shuai Wang, and Bing Liu. Learning cumulatively to become more knowledgeable. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1565–1574, 2016. ISBN 978-1-4503-4232-2.

Óscar Fontenla-Romero, Bertha Guijarro-Berdiñas, David Martinez-Rego, Beatriz Pérez-Sánchez, and Diego Peteiro-Barral. Online machine learning. *Efficiency and Scalability Methods for Computational Intellect*, pages 27–54, 2013.

Artur d'Avila Garcez, Tarek R. Besold, Luc de Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C. Lamb, Risto Miikkulainen, and Daniel L. Silver. Neural-Symbolic Learning and Reasoning: Contributions and Challenges. *Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, Stanford*, 2015.

John Haugeland. *Artificial Intelligence: The Very Idea*. Cambridge, Mass: MIT Press, 1985.

John E. Laird and Robert E. Wray III. Cognitive architecture requirements for achieving AGI. In *Proc. of the Third Conference on Artificial General Intelligence*, pages 79–84, 2010.

S. Legg and M. Hutter. A collection of definitions of intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, 157(2):17–24, 2020.

John McCarthy. Artificial intelligence needs more emphasis on basic research: President's quarterly message. In *AI Magazine*, volume 4, page 5, 1983.

Marvin Minsky. *Society of mind*. Simon and Schuster, 1988.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. *Commun. ACM*, 61(5):103–115, 2018. ISSN 0001-0782.

James Moor. The dartmouth college artificial intelligence conference: The next fifty years. In *AI Magazine*, volume 27, pages 87–91, 2006.

Erik T. Mueller and Henry Minsky. Using thought-provoking children's questions to drive artificial intelligence research. *CoRR*, abs/1508.06924, 2015. URL http://arxiv.org/abs/1508.06924.

Allen Newell. *Unified theories of cognition*. Harvard University Press, Cambridge, MA, 1994.

Nils J. Nilsson. The physical symbol system hypothesis: Status and prospects. In *M. Lungarella et al. (Eds.), 50 Years of AI, Festschrift, LNAI 4850*, pages 9–17. 2007.

E. Nivel, K. R. Thórisson, B. R. Steunebrink, H. Dindo, G. Pezzulo, M. Rodriguez, C. Hernandez, D. Ognibene, J. Schmidhuber, R. Sanz, Helgi Páll Helgason, A. Chella, and G. K. Jonsson. Bounded Recursive Self-Improvement. Technical RUTR-SCS13006, Reykjavik University Department of Computer Science, Reykjavik, Iceland, 2013.

Eric Nivel and Kristinn R. Thórisson. Towards a Programming Paradigm for Control Systems with High Levels of Existential Autonomy. In Kai-Uwe Kühnberger, Sebastian Rudolph, and Pei Wang, editors, *Proceedings of the Sixth Conference on Artificial General Intelligence*, volume 7999 of *Lecture Notes in Artificial Intelligence*. Springer, Beijing, China, 2013.

K. Panton, C. Matuszek, D. Lenat, D. Schneider, and et al. Common sense reasoning - from Cyc to intelligent assistant. *Ambient Intelligence in Everyday Life*, 2022.

J. Pearl. *do*-calculus revisited. In Nando de Freitas and Kevin Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 4–11, Corvallis, OR, 2012. AUAI Press.

Jean Piaget. *The construction of reality in the child*. Routledge, 2013.

John L. Pollock. Defeasible reasoning and degrees of justification. *Argument and Computation*, 1(1):7–22, 2010.

Tiernan Ray. AI's true goal may no longer be intelligence. *ZDNet.com,*, 2022.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

T.D. Sanger. Neural network learning control of robot manipulators using gradually increasing task difficulty. *IEEE Transactions on Robotics and Automation*, 10(3):323–333, June 1994. ISSN 1042-296X. doi: 10.1109/70.294207.

Rylan Schaeffer, Mikail Khona, and Ila Rani Fiete. No free lunch from deep learning in neuroscience: a case study through models of the entorhinal-hippocampal circuit. In *Neural Information Processing Systems*, 2022. doi: https://doi.org/10.1101/2022.08.07.503109.

Arash Sheikhlar, Kristinn R. Thórisson, and Leonard M. Eberding. Autonomous cumulative transfer learning. In *International Conference on Artificial General Intelligence*, pages 306–316. Springer, 2020.

Arash Sheikhlar, Leonard M Eberding, and Kristinn R. Thórisson. Causal generalization in autonomous learning controllers. In *International Conference on Artificial General Intelligence*, pages 228–238. Springer, 2021.

Daniel L. Silver, Qiang Yang, and Lianghao Li. Lifelong Machine Learning Systems: Beyond Learning Algorithms. In *AAAI Spring Symposium: Lifelong Machine Learning*, 2013.

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 229(October), 2021. doi: https://doi.org/10.1016/j.artint.2021.103535.

Herbert A. Simon. Artificial intelligence: An empirical science. *Artificial Intelligence*, 77: 95–127, 1995.

Aaron SlomanAaron Sloman. Architectural requirements for human-like agents, both natural and artificial: what sorts of machines can love? In *Human Cognition and Social Agent Technology*, pages 163–190. John Benjamins Publishing Company, Amsterdam, NL, 2000.

Jerry Swan, Eric Nivel, Neel Kant, Jules Hedges, Timothy Atkinson, and Bas Steunebrink. *The Road to General Intelligence*. Springer Studies in Computational Intelligence, volume 1049, 2022. doi: https://doi.org/10.1007/978-3-031-08020-3.

Kristinn Thórisson and Helgi Helgason. Cognitive Architectures and Autonomy: A Comparative Review. *Journal of Artificial General Intelligence*, 3(2):1–30, January 2012. ISSN 1946-0163. doi: 10.2478/v10229-011-0015-3.

Kristinn R. Thórisson. A new constructivist AI: from manual methods to self-constructive systems. In *Theoretical Foundations of Artificial General Intelligence*, pages 145–171. Springer, 2012a.

Kristinn R. Thórisson. A New Constructivist AI: From Manual Methods to Self-Constructive Systems. In Pei Wang and Ben Goertzel, editors, *Theoretical Foundations of Artificial General Intelligence*, volume 4 of *Atlantis Thinking Machines*, pages 145–171. Atlantis Press, Amsterdam, The Netherlands, 2012b.

Kristinn R. Thórisson. Reductio ad Absurdum: On Oversimplification in Computer Science and its Pernicious Effect on Artificial Intelligence Research. In *Proceedings of the Sixth Conference on Artificial General Intelligence*, volume 7999 of *Lecture Notes in Computer Science*, Beijing, China, 2013. Springer.

Kristinn R. Thórisson. Machines with Autonomy & General Intelligence: Which Methodology? In *Proceedings of the Workshop on Architectures for Generality and Autonomy 2017*, Melbourne, Australia, August 2017.

Kristinn R. Thórisson. Seed-programmed autonomous general learning. In *Proceedings of Machine Learning Research*, volume 131, pages 32–70, 2020.

Kristinn R. Thórisson. The explanation hypothesis in autonomous general learning. In *Proceedings of Machine Learning Research*, volume 159, pages 5–27. Springer, 2021.

Kristinn R. Thórisson and David Kremelberg. Understanding and common sense: Two sides of the same coin? In *Proc. Artificial General Intelligence*, pages 201–211, 2018.

Kristinn R. Thórisson and Arthur Talbot. Cumulative learning with causal-relational models. In *Artificial General Intelligence*, pages 227–237, Cham, 2018. Springer International Publishing.

Kristinn R. Thórisson, Eric Nivel, Bas R. Steunebrink, Helgi Páll Helgason, Giovanni Pezzulo, Ricardo Sanz, Jürgen Schmidhuber, Haris Dindo, Manuel Rodriguez, Antonio Chella, Gudberg K. Jonsson, Dmitri Ognibene, and Carlos Hernandez. Autonomous Acquisition of Situated Natural Communication. *Computer Science & Information Systems*, 9(2):115–131, 2014. Outstanding Paper Award.

Kristinn R. Thórisson, David Kremelberg, Bas R. Steunebrink, and Eric Nivel. About understanding. In *Proceedings of AGI-16*, pages 106–117, New York, NY, USA, 2016. Springer-Verlag.

Kristinn R. Thórisson, Jordi Bieger, Xiang Li, and Pei Wang. Cumulative learning. In *International Conference on Artificial General Intelligence*, pages 198–208. Springer, 2019.

David Waltz. The importance of importance: Aaai-98 presidential address. In *AI Magazine*, volume 20, pages 18–35, 1999.

P. Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10 (2):1–37, 2020.

Pei Wang. *Rigid Flexibility.* Springer, 2006.

Pei Wang. Insufficient Knowledge and Resources-A Biological Constraint and Its Functional Implications. In *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures.* Citeseer, 2009.

Yusen Zhan and Matthew E. Taylor. Online Transfer Learning in Reinforcement Learning Domains. *arXiv preprint arXiv:1507.00436*, 2015.

Du Zhang. From one-off machine learning to perpetual learning: A step perspective. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.