WHAT IS INTELLIGENCE IN THE CONTEXT OF AGI?
 dbruiger [at] gmail.com

Abstract

Lack of coherence in concepts of intelligence has implications for artificial intelligence. 'Intelligence' is an abstraction grounded in human experience while supposedly freed from the embodiment that is the basis of that experience. In addition to physical instantiation, embodiment is a condition of dependency, of an autopoietic system upon an environment, which thus matters to the system itself. The autonomy and general capability sought in artificial general intelligence implies artificially re-creating the organism's natural condition of embodiment. That may not be feasible; and, if feasible, it may not be controllable or advantageous.

1. Intelligence in the eye of the beholder?

How coherent is the concept of 'intelligence'? It remains, after all, an ill-defined and still controversial notion, now used roughly to mean 'goal directed adaptive behavior' [1]. Intelligence has also been defined as the ability to learn, to deal with novel situations, to adapt with insufficient information, to do abstract thinking, etc. It has even been defined as the ability to score well on intelligence tests! It is commonly accepted in AI that "intelligence measures an agent's ability to achieve goals in a wide range of environments [2]." Such a statement glosses over divergent possible understandings of goals, environments, and agents. An ambiguity common to many definitions is that intelligence can refer either to observed *behavior* or to an inner *capacity* or potential. It is tempting, if circular, to reify observed behavior as a property or capacity inferred to be responsible for that behavior [3]. Little more is thereby gained than by Moliere's "dormitive principle."[1] Intelligence is often misconceived as somehow substantial, which one philosopher wryly calls *smartonium* [4].

It may be equally ill-advised to conflate various behaviors or skills into one overall capacity, such as "problem-solving ability" or Spearman's famous *g*-factor. While ability to solve one sort of problem carries over to some extent to other sorts of tasks, it does not necessarily transfer equally well to all tasks, let alone to activities that might not best be described as problem solving at all—such as, for example, the ability to be happy. Moreover, problem *solving* is a different skill from finding, setting, or effectively defining the problems worth solving. The challenges facing society usually seem foisted upon us by external reality, often as emergencies. Our default responses and strategies are often more defensive than proactive. Another level of intelligence might involve better foresight and planning. Concepts of intelligence may change as our environment becomes progressively less natural and more artificial, consisting largely of other humans and their intelligent machines.

Like the notion of mind-in-general, intelligence is an abstraction that is grounded in human experience while theoretically freed from the embodiment that is the basis of that experience. It is understandably anthropocentric, derived historically from comparisons among human beings, and then extended to comparisons of other creatures with each other and with human beings. In AI, the goal is to produce machines that can behave "intelligently"—in some abstract sense that is

[1] In which a sleeping potion is said to be effective because of the 'dormitive principle' it contains.

extrapolated from biological and human origins. This paper explores the limits and coherence of what that might mean.[2]

Here we propose that the context and ultimate referent for a meaningful concept of intelligence is necessarily biological. Biologically, intelligence is the ability to survive. All living things are tautologically successful, therefore intelligent, given that the Aristotelian "final" goal of life is its own continuance. Though trivial sounding, this is important to note because models of intelligence are grounded in experience with organisms and because the ideal of artificial general intelligence (AGI) involves attempting to create artificial organisms that are, paradoxically, freed of the constraints of biology. If intelligence is neither inherently anthropocentric, nor an embodied product of selection, there may be essentially no consistent basis for a definition and no constraints on what intelligence may consist of other than those imposed by physical laws. In other words, a reasonable definition of intelligence must be grounded in biology. Alien intelligence, for example, need not resemble human intelligence any more than the alien creature need resemble the human form. Nevertheless, we must presume that any living creature would somehow be a product of natural selection. The ideal of AGI appears deceptively free of that constraint; yet, it may turn out that the only way for an AI to have the desired autonomy and generality is to be a product of some form of selection. A relevant point is that, if an AI does not constitute an artificial organism, then its intelligence is not actually its own but that of its creators. While it appears that autonomy is a question of degree, there is a categorical difference between a truly autonomous *agent* and a mere *tool*. It is relevant to ask *whose* intelligence is involved, since a tool manifests only the derived intelligence of the agent designing or using it.

An organism is an agent in the sense that it acts on its own behalf, originating action with its own self-renewing energy and for its own purposes—which primarily concern its own well-being or that of its kind. In other words, it is an adaptive *autopoietic system*: one whose chief product is itself [5]. A species can adapt through natural selection, in which less than optimal solutions are weeded out. Individual specimens can adapt through self-modification, in which less optimal solutions (however they are stored or represented) are weeded out through ongoing experience. Posing problems or challenges to an agent (for example, to test or train its intelligence) is like presenting a game. The presumed object is to win, to achieve the goal defined by the tester. But the task from the point of view of the agent may be to assimilate the game and its rules to the agent's own goals. The situation is different according to whether the challenge (test/training) is posed by the natural environment, by another agent, or by the agent itself. Tests and machine learning situations are artificial. The ultimate test and learning situation for organisms is the challenge to live.

Many behaviors in organisms that seem to be the product of reasoning or learning are actually inflexible adaptations—that is, reflexes or instincts [6]. Insofar as they serve the organism, they are no less intelligent for that. What then is the difference between teleology and causality, when they seem to use the same physical processes? While the question might be spurious, it does point to something essential: the observer's relation to the phenomena observed or created in laboratory. In the name of objectivity, science generally excludes the subject's or observer's role. Thus, concepts such as intelligence, agency, intentionality, knowledge, fact, information, truth—and even teleology—are artificially divorced from their dependency on a subject.[3] Information, for example, can be defined as the reduction of uncertainty. But it is always some*one*'s uncertainty that is

[2] If intelligence is understood as the ability to think abstractly, then the notion of intelligence is itself intelligent. That conclusion is compromised, however, to the degree the concept is not self-consistent.

[3] E.g., the assumption that every proposition simply *is* either true or false ignores the intervention of a subject's belief, uncertainty, action to test or prove the proposition, discovery of intermediate or indeterminate cases, etc. Causality and teleology are assumed to be natural concomitant of 3rd-person (space-time) description, but (like space and time) are categories invented by the brain.

reduced. Despite Shannon, information is not an absolute property of a system. There might be commonality and agreement among subjects, but it is not as though no subject is involved. Similarly, "intelligence" is someone's intelligence, measured *by* someone with a particular yardstick for particular purposes.

2. General intelligence

Measures of intelligence were developed to evaluate human performance in various areas of interest to the measurers. This gave rise to a notion of general intelligence that could underlie specific abilities. A hierarchical concept of intelligence would have a "domain independent" skill (*g*-factor) that informs and perhaps controls domain-specific skills. "General" can refer to the range of situations and also to the range of subjects. What is general across humans is not the same as what is general across known species or theoretically possible agents or environments. Following Ashby's Law of Requisite Variety, the intelligence measured can be no more general than the tests used to measure it.[4]

It is difficult to compare animal intelligence across species, since wide-ranging sense modalities, cognitive capacities, and adaptations are involved. Tests may be biased by human motivations and sensory-motor capabilities; the tasks and rewards for testing animal intelligence are defined by humans, aligned with their goals. Even in the case of testing human beings, despite wide acceptance and appeal,[5] Spearman's *g*-factor has been criticized as little more than a reification whose sole evidence consists in the behaviors and correlations it is supposed to explain [7]. Nevertheless, the comparative notion of intelligence, generalized across humans, was further generalized to include other creatures in the comparison (G-factor for the species). The concept of *artificial* general intelligence generalizes this even further to include machines. Since it should not be anthropocentric—and should be independent of particular sense modalities, environments, goals, and even hardware—Legg and Hutter propose a concept of *universal* intelligence that can apply to "arbitrary systems." They even suggest the possibility of a universal test [8].[6]

In organisms, however, the evolution of specific adaptive skills must be distinguished from the evolution of a general skill called intelligence. In conditions of relative stability, natural selection would favor automatic domain-specific behavior, reliable and efficient in its context. Any pressure favoring *general* intelligence would arise rather in unstable conditions. The emergence of domain-general cognitive processes would translate less directly into fitness-enhancing behavior, and would require large amounts of costly brain tissue. The question for evolutionary theory is how domain-general processes could evolve "on top of" domain-specific adaptations and what would drive their emergence [9]. These are questions relevant to AGI. The circumstances that lead to an increase in

---

[4] If intelligence is defined as the ability to pass intelligence tests, then a specific test could be customized to defeat any given agent. In other words, there cannot be a fixed test (or algorithm) for general intelligence.

[5] For instance, *"...g* in humans has a clear genetic foundation … Furthermore, *g* has robust correlates in brain structure and function, such as brain size, gray matter substance, cortical thickness, or processing … Finally, *g* is also a good predictor for various measures of life outcome, including school achievement, the probability of being in professional careers, occupational attainment, job performance, social mobility, and even health and survival." [Burkhardt et al, p5]

[6] "…if we could formally define and measure the complexity of test problems using complexity theory we could construct a formal test of intelligence. The possibility of doing this was perhaps first suggested by Chaitin... While this path requires numerous difficulties to be dealt with, we believe that it is the most natural and offers many advantages: It is formally motivated, precisely defined and potentially could be used to measure the performance of both computers and biological systems on the same scale without the problem of bias towards any particular species or culture." [Legg & Hutter, p36]

natural general intelligence—rather than (or in addition to) specific skills—may shed light on the differences between an AI *agent* or *tool*.

In light of the benefits of general intelligence, why do not all species evolve bigger and more powerful brains? Every living species is by definition smart enough for its current niche [10]. Its intelligence is an economical adaptation to that niche. It would seem, as far as life is concerned, that general intelligence is not only costly but implies a general niche, whatever that can mean. Humans, for example, evolved to fit a wide range of changing conditions and environments, which they continue to further expand through technology. Even as (or if) we stabilize the natural environment, the human world changes ever more rapidly—requiring more general intelligence to adapt to it. "Universal" intelligence would imply a universal environment—that is, one so abstract that it is completely liberated from all particulars, including human conceptions! Yet, universal intelligence is but a human idea and ideal, constrained to the limits of imagination and the world as presently conceived. We do, of course, recognize such limits and that the future is unknown. It is perhaps in order to reassure ourselves in the face of that uncertainty that we conceive of intelligence somewhat as a magical power to deal with the unknown.

A general intelligence factor at the head of a hierarchy of skills could involve executive functions, such as inhibitory controls and selective attention—in other words, conscious control [11]. Yet, a definition of the intelligence or cognition of an agent need not involve phenomenality,[7] let alone self-consciousness. Because the only cognition we know subjectively is our own, we tend nevertheless to think of it as involving some conscious content (if not our own, then something nebulously like it or some "degree" of it). But cognition and intelligence can both be regarded in purely behavioral terms. And the primary behavior involved for artificial agents (as for organisms) is self-production (autopoiesis), which does not of itself imply consciousness. While personhood presumes agency, and goes beyond it, agency itself can be considered from the 3rd-person. Even valuation does not have to imply conscious sensation or feeling, though for humans it obviously involves them. On the other hand, while concepts of general intelligence are based on human experience and performance, it remains unclear to what extent an AI could satisfy the criteria for human-level general intelligence without itself being—if not conscious—at least an embodied autonomous entity: effectively an organism.

Creating programs for computers diverges from creating biological-like robots. The former tends to aim for general intelligence capable of solving human mental challenges; the latter tends toward simple artificial creatures that imitate the physical behavior of simple natural creatures. The former approaches from the top down, the latter from the bottom up. But neither computer simulations of embodiment nor physical robots constitute embodied agency as here understood—so far.

3. Intelligence as computation

The possibility to understand mind as computation, and to view the brain metaphorically as a computer, is one of the great achievements of the computer age. Computer science and brain science have productively cross-pollinated. Yet, *mind* is a vague concept not exclusively related to brain. Mind and thinking suggest reasoning and an algorithmic approach—the ideal of intellectual thought—which is only a small part of the brain's activity responsible for the organism as a whole. The computer metaphor is underwritten more broadly by the mechanist metaphor, which holds that *any* behavior of a biological "system" could be reduced to an algorithm. However, the idea of *system* is already an idealization in which the structure and behavior of an organism is redefined in mechanist terms. (Being *defined* in the first place, it is reconfigured to consist only of defined parts

[7] I.e., qualia

and relationships.) The natural thing is no longer a product or denizen of nature, but now a product of formal definition and a denizen of a human conceptual domain. The ability to define, to abstract and idealize, and to manipulate such abstractions in a thought world, is considered the hallmark of intelligence. Yet the degree to which idealizations fail to fit reality, with possible serious consequences, calls that intelligence into question.

The converse of the challenge to understand natural intelligence through computation is to understand machine intelligence in terms of (human) thought processes. Part of that quest should be to understand how our own mental processes—including language—unconsciously shape concepts of machine intelligence. It seems reasonable that to *understand* the lawful (i.e., algorithmic) basis of behavior in organisms, one should attempt to program a computer to *perform* that behavior: "to build intelligence" rather than simply to study the external behavior of intelligent agents or dissect their inner workings [12]. However, "that behavior" already presumes a description—that is, a reduction of the behavior to a linguistic expression. Creating a program to produce action that meets that description does not exhaust the reality of what the organism is doing. The catch is that the behavior as described in language and as prescribed in the algorithm may correspond to each other (being reducible to a common formula), while neither corresponds perfectly to the real behavior of the organism, which is *not* a literal machine nor a matter of human definition. This poses a serious limitation on the computational metaphor. It also shows how language, and categories of thought based upon it, can mislead. To conceive of "flight" as an abstraction not pinned to creatures flapping their wings has enabled us to create machines that "fly." While aerodynamics seems to liberate flying from biology, the common basis between birds and airplanes is not only aerodynamics; it is also the intent to move freely in 3 dimensions, to get from one place to another, etc. These are creaturely motives and human creatures are not independent them. If we view intelligence in that light, it is grounded in biological needs as much as flight is. Similarly, a human baseball pitcher performs "the same" action as a baseball pitching machine. But the devil is in the details, which are obscured by the rubric "pitching." The rubrics "thinking" and "intelligence" may obscure important details and mislead us when trying to assimilate within a single concept the activity of machines and of brains.

4. Natural intelligence is embodied

Physical instantiation is a necessary but not sufficient condition for *embodiment*, which further requires a relationship of dependency upon an environment. The precarious dependency of life on environment is the reason why things *matter* to organisms and not to machines [13]. True (i.e., "strong") embodiment for a robot would thus mean that it is integrated and connected to the world in the same ways as an organism: not only through a sensory-motor interface but also through its own valuations, of events that matter to it because of its critical dependency on the world. For organisms, this state of affairs comes about through natural selection, such that only those systems exist that are internally organized to behave in ways that preserve, or at least permit, their continuing existence. An autopoietic system is a type of homeostatic system that tries to keep constant the conditions required for its own existence. The "intelligence" of the autopoietic system is ultimately its ability to maintain and preserve itself.

Robots resemble organisms insofar as their hardware and software are integrated in a physical body. On the other hand, AI is usually not physically instantiated. It consists of software in a computer, which may, however, have access to the real world through sensors and controls over real systems. To be an *agent*, it must be embodied in the further sense of being an autopoietic system. That means it will have its own intelligence and goals, in pursuit of its own existence, which may conflict with the human programmer's goals.

Much of the contemporary study of "embodied" systems focuses on the effects of the particular physical instantiation of robots on locomotion and cognition—especially on how morphology itself can substitute for representation. However, this ignores the crucial relevance of embodiment as an autopoietic relationship with the world. It tends also to be limited to relatively simple creatures and their robotic versions. Yet, the research commitment to autonomy and general intelligence in pursuit of total automation leads inexorably in the direction of true embodiment. Paradoxically, that means that the desire to extend control indefinitely through automation may lead ultimately to a *loss* of control—because embodied AI will have a will of its own. Of course, quite apart from presumed human utility, another motive to re-create embodiment artificially is the sheer challenge to match nature by creating artificial life. This may have the added benefit of deeper understanding, on Vico's principle that we truly understand only that which we make. The goal is not to *simulate* intelligence so much as to actually synthesize it in real systems [14]—if that is even a valid distinction in this context. And if, indeed, the benefits of intellectual understanding outweigh the risks of such a venture.

Whatever its motivation, recreating embodiment may be easier said than done. Nature effected it through natural selection over generations of reproducing and dying individuals. Artificial evolution of software takes place within computers and does not involve physical instantiation. The evolved software can be physically connected to hardware after the fact, but this is not the same as natural evolution any more than brains (or, for that matter, genes) evolve separately from their bodies. The unit of selection is the physical individual, which is an integration of software and hardware in a body that self-produces in the real world. A physical robot has that integration but is not self-producing, let alone reproducing. The artificial equivalent of natural selection for robots would imply generations of robot bodies wastefully (and painfully) destroyed.

While it is conceivable for a robot body to self-assemble, self-maintain, and adaptively reconfigure itself, a key difference between organisms and machines is that creatures consist of parts (cells) that are autopoeitic on their own scale, which means they too can reconfigure themselves adaptively. A machine might be self-repairing, but might not consist of parts that are self-repairing. A machine might conceivably have its own goals (related to its own well-being), but not consist of parts that have *their* own goals, subordinated somehow to the well-being of the machine as a whole. To re-create the agency of organisms artificially, it might be necessary to re-create the composite structure of organisms: machine components (as well as computational ones) that can autonomously and adaptively reconfigure themselves physically and also put themselves together in just the right way [15]. This is a matter that deserves further investigation, at the intersection of biology and computation. A further consideration, often overlooked, is that autopoietic systems are not only self-producing and self-repairing but also self-defining. In the case of AI, that would mean self-programming. An individual organism inherits an initial genetic program; yet, on a species level, that program was also self-producing. Its unfolding also depends on a multitude of "epigenetic" factors. This raises the question: how, and to what extent, can human programmers stand in for millions of years of natural self-programming, even to provide a minimal "seed" of a self-developing AI? How much of that can be hard-coded initial information versus some source of guidance for further learning and growth [16]?

Must an AI be an agent to achieve full autonomy, with human-level general intelligence and beyond? If so, the agent itself decides how to divide up the world and to reason about it [17]. But on what basis, if not in terms of its own preferences? Learning about the preferences of others is not the same as having those preferences oneself. Training of animals is done with rewards, but the reward is valued by the animal because it already has needs and preferences of its own. Training of AI cannot mean the same thing unless the AI is similarly an agent. Achieving a goal makes no difference to the AI unless it is an autopoietic system. Rewards would make no difference unless

the system "values" them, and value can only come from the consequence to a system that matters to the system itself. How does a machine come to have preferences of its own, to *value* something at all? Short of that, the machine can only do what it is told to do, even when that includes learning from its interpretation of human preferences and behavior.

While it is true that even neural networks can be fooled because the situation they are in means nothing to them [18], it hardly follows that agency would increase reliability from a human standpoint. Artificial agents may be unreliable because they have a will of their own; but also because they may be vulnerable to the kinds of error that human agents are and for similar reasons. Reliability toward humanly-defined goals can be achieved in AI simply by adding refinements ad hoc to non-agential tools, asymptotically reducing the chance of error. The system—such as the self-driving car—need not be fool-proof; it need only reach a desired statistical reliability. In other words, in many situations there may be no net advantage to creating an agent. The ideal of intelligence as a magic bullet, tacitly behind AGI, refers to the ability of an agent to look after itself, not its capacity to satisfy humans.

Abilities—of people, animals, or machines—are human concepts. The currency of "smartonium" is ultimately backed up by success in a real environment. Not only can adaptability be a function of brain size, but the converse may be true insofar as survival rate is a precondition for the opportunity to evolve larger brain size. Only species not subject to high predation pressure would have that opportunity [19]. Brain size and intelligence for primates are a function as well of opportunities for social learning; the intelligence of our species correlates with human socialization, including intense maternal care. Yet, the relationship between general intelligence and socio-cognitive abilities in humans remains incompletely understood [20]. Socialization for artificial intelligence has scarcely been studied. And the incompletely understood role of consciousness in general intelligence leaves unanswered the question of whether the high expectations of AGI could be fulfilled without some equivalent of human consciousness.

5. Whose agency?

The tasks of the conscious human person are not (necessarily) the tasks of the biological human organism. It is probably this psychological fact that allows us to conceive of an ideal agent pursuing arbitrary goals, whereas the goals of living things are hardly arbitrary. As far as it is relevant to humans, the intelligence of other entities (whether natural or artificial) must be measured by their capacity to further or thwart human aims. Whatever does not interact with us in ways of interest to us may not even be recognized as intelligence.

An AI can be a goal-seeking artifact—such as a guided missile, robot, expert system, or infobot—without being an agent in the sense explored here. In particular, such artifacts are allopoietic systems, which take for granted—or leave out of the picture—exactly *whose* goals are pursued and why. Perhaps this omission arises because agency in general is out of bounds for the scientific version of naive realism, which excludes any form of subjectivity [21].[8] A similar problem besets biology, when the organism is considered only from the point of view of the observer yet the challenge is to understand how things can be meaningful for a system from its own perspective [22]. An organism is not only an object for an observer, but is also a subject in its own right—not necessarily in the sense of having phenomenal experience, but at least in the sense of being an agent, acting in its own interests.

[8] It also dodges epistemic, ethical, and legal responsibility.

Information is "a difference that makes a difference" [23] *to an agent*. A difference in some domain must be recognized by a cognitive system that maps that difference to some difference within itself, upon which it can act according to an agenda that favors or at least permits the system's continuing existence. Only then is it information *for* the system. If the difference matters only to the observer, then it is information for the observer but not for the system in question; if it matters only to the system, it will likely not even be noticed by the observer. That is a key difference between autopoietic and allopoietic systems: the information processed by an AI *agent* matters to it and is used by it for its own purposes, whereas the information processed by an AI *tool* (such as a computer) exists only for the mind and purposes of the human user. A glib use of language permits speaking of programming AI tools to "have" goals; but clearly the goals are the programmer's. Similarly, the intelligence of the AI tool is the intelligence of the programmer. What does it mean, then, to create AI that is more intelligent than humans? Straightforwardly, it can mean that a skill valued by humans is automated to more effectively achieve human goals. We are used to this idea, since every tool and machine was motivated by such improvement and usually achieves it until something better comes along. But is *general* intelligence a skill that can be so augmented, automated, and treated as a tool? Or are only agents capable of it? If the latter, then their intelligence will serve their own interests, with no guarantee how these will overlap with human interests.

6. Motivation

This brings us to the question: why pursue AGI? If the goal is to build powerful and flexible machines [24] to serve human purposes, then we should shy away from the chimeric idea that agents can be made to serve human purposes better than tools. Whether or not they have a subjective inner life, they will be dedicated to their own purposes.[9] We tried augmenting human labor with animal slaves and then built machines as more efficient and powerful replacements. We tried human slavery as well. But a machine designed to replace the human slave—by a service cheaper, more reliable, and even more intelligent—is the goal of the designer, not of the machine. However devoted to the master, the goal of a human slave is to live. This would be no less true of artificial agents. Quite apart from ethics, we cannot expect intelligent artificial slaves to remain subservient. To create artificial agents for some *other* reason—for example, as companions or mentors—would follow an entirely different motivation. In that case, such agents might, like us, seek to live and work together harmoniously as part of an extended society [25]. Given the example of human history, however, one shouldn't count on it.

7. Superintelligence

Even if speed were the only criterion, it might seem reasonable to suppose that any electronic equivalent of a natural organism would at least be as many times more "intelligent" as the speed of electricity is faster than electrochemical processes. Yet, the idea of superintelligence can be no more coherent than our concepts of "ordinary" intelligence. (In fact, far less so, since we might not be able even to comprehend it.) If universal intelligence (the *u*-factor?) is not a coherent notion, then neither is *u*+. If smartonium is a myth, then so is supersmartonium. While there is evidence for the *g*-factor across human beings, the prospect of automating and amplifying it in AI is a different and questionable matter.

---

[9] Subjective qualities such as understanding, consciousness, creativity, imagination, free will, emotion, etc. "are only relevant to our goal to the extent to which they have some measurable effect on performance in some well-defined environment… The question is whether they are relevant or not." [Legg & Hutter p42]

What do we expect of artificial superintelligence? *We want it to do what we want, better than we can, and without supervision.* This raises several questions—and should raise eyebrows. *Will* it do what we want, or how can it be made to do so? How will we trust its (superior) judgment if its considerations are unintelligible to us? How autonomous can AI be short of being an agent? Can a sophisticated non-agential AI be accidentally triggered to become an agent? What are the necessary conditions for AI to become an agent?

Other questions arise as well. Motives, desires, and goals have evolved in organisms through natural selection; animals have been bred and trained to fulfil human wishes. Can goals or motives be *programmed* into AI agents, either built in from scratch or trained in through machine learning? How will that differ from programming or training AI tools? Superintelligent artificial agents, if such are feasible, may pose a threat to human beings whether or not we consider them conscious. But what about superintelligent tools? Is an AI takeover by non-agential tools a threat and what would it look like?

8. Conclusion

Simon and Newell's founding premise for AI is that formal symbol manipulation is both a necessary and sufficient mechanism for general intelligent behavior. The weakness of this thesis is that "general" intelligent behavior is less general than supposed, referring (circularly?) to the restricted sorts of behavior that can be formalized in order to accomplish specific kinds of goals. The *ideal* is to be able to inform a flexible AI of *any* goal and rely on it to efficiently achieve it. The "intelligence" presumed for this degree of flexibility may, however, imply a system that has its own competing goals: an agent, which is an adapting autopoietic system. The ideal of universal intelligence implies a purely syntactic computerized version of the general intelligence manifested by living things (humans, in particular). It extends the ideal of the universal machine, in the form of a powerful system that can accept and carry out any arbitrary goal but has no goal of its own. This may prove to be a contradiction in terms. One tacitly expects intelligence to be fully agential, with semantic and real-world interaction, while inconsistently hoping for a non-agential version—or, at least, a subservient slave. Indeed, slavery illustrates the problem of control and the dilemma of programming agents externally.

While concepts of intelligence are grounded in human experience (consciousness), they have been abstracted to constitute a property theoretically independent both of consciousness and embodiment, and transferable to non-living systems. However, the ideal of universal intelligence fails to be liberated from biology and anthropocentrism. In contrast to its disembodied view of cognition, a "biogenic" view (as opposed to an "anthropogenic" one) emphasizes the organism's ongoing assessment of its situation and its own valuation of stimuli in terms of its own embodied well-being [26]. The quest for superintelligence raises the question: under what circumstance could machines become self-interested agents, competing with each other and with humans and other life forms for resources and their very existence? The dangers of superintelligence attend the motive to achieve ever greater autonomy, the extreme of which is the genuine autonomy manifest by living things. AI should instead focus on creating powerful tools that remain under human control. That would be safer, wiser—and shall we say more intelligent?

REFERENCES

[1] R.J. Sternberg and W. Salter "Conceptions of Intelligence" in *Handbook of Human Intelligence* ed by R.J. Sternberg Cambridge UP 1982, p3

[2] Shane Legg & Marcus Hutter “Universal Intelligence: a definition of machine intelligence” 2007, arXiv:0712.3329v1, p12

[3] Henry D. Schlinger “The Myth of Intelligence” The Psychological Record, 2003, 53, 15-32

[4] Thomas G. Dietterich “How to create an intelligence explosion—and how to prevent one” *Edge:* What do you think about machines that think? June 2018

[5] Humberto Maturana and Francisco Varela *Autopoiesis and Cognition* Reidel, 1980

[6] David Poole, Alan Mackworth, and Randy Goebel *Computational Intelligence: A Logical Approach* Oxford University Press, New York. 1998, Chapter One, p7

[7] Schlinger, op cit

[8] Legg & Hutter, op cit, p4

[9] J. M. Burkhardt, M. N. Schubiger, C. P. van Schaik “The Evolution of General Intelligence” 2017, Zurich Open Repository and Archive, p3-10

[10] ibid, p35

[11] ibid, p6-7

[12] Poole et al, op cit, p7

[13] Jonas, H. ([1966] 2001). *The Phenomenon of Life: Toward a Philosophical Biology*. Northwestern University Press

[14] Poole et al, op cit, p2

[15] Fitch, W. T. (2008). "Nano-intentionality: A defense of intrinsic intentionality," *Biology and Philosophy* 23: 157-177, Sec 6

[16] Mark R. Waser “What Is Artificial General Intelligence? Clarifying the Goal for Engineering and Evaluation”, 2008

[17] Poole et al, op cit, p11

[18] Tom Froese and Shigeru Taguchi “The Problem of Meaning in AI and Robotics: Still with Us after All These Years” *Philosophies* 2019, *4*, 14, sec2

[19] Burkhardt et al, op cit, p35

[20] ibid

[21] Dan Bruiger *The Found and the Made,* Transaction/Routledge, 2016

[22] Tom Froese “Life is precious because it is precarious: Individuality, mortality, and the problem of meaning” in G. Dodig-Crnkovic and R. Giovagnoli (Eds.), *Representation and Reality in Humans, Animals and Machines*, Springer, 2017

[23] Gregory Bateson *Steps to an Ecology of Mind* U. of Chicago Press, 1972

[24] Legg & Hutter, op cit, p42

[25] Waser**,** op cit

[26] Pamela Lyon “The biogenic approach to cognition” Cogn Process (2006) 7: 11–29