Advances in Cognitive Systems 9 (2020) 1–6 Submitted 5/2020; published 7/2020

General Intelligence Requires Autonomous, Cognitive, Intentional Agents

Sean Kugele SEANKUGELE@GMAIL.COM
Stan Franklin FRANKLIN.STAN@GMAIL.COM
Department of Computer Science and Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152 USA

Abstract

We contend that natural and artificial systems exhibiting "general intelligence" will likely be endowed with three properties. First, they must be *autonomous*; that is, capable of acting in pursuit of their own agendas. Second, they must be *cognitive*; that is, capable of reflecting on, and reasoning about, their environments in a manner that is decoupled from their immediate inputs and outputs. Third, they must be *intentional*; that is, capable of connecting the content of their mental states to corresponding referents in their environments. We argue that deficits in any one of these properties can lead to pathological behavior in humans and other animals, and that the complete absence of any of these is incompatible with general intelligence.

1. Introduction

Newell and Simon (1976) characterized "general intelligence" as the ability to perform actions that show the same "scope of intelligence" as human actions, are "appropriate to the ends of the system," and are "adaptive to the demands of the environment… within some limits of speed and complexity." Goertzel and Pennachin (2007) stated that "[a] general intelligence must be able to carry out a variety of different tasks in a variety of different contexts, generalizing knowledge from one context to another, and building up a context and task independent pragmatic understanding of itself and the world." And Voss (2007) stated that "[t]he mark of a *generally* intelligent system is not *having* a lot of knowledge and skills, but being able to *acquire* and *improve them*—and to be able to appropriately *apply* them." While all of these notions are broadly consistent, they focus on different aspects of general intelligence, and say little about the characteristics of the "control structures" (Newell, 1973) from which generally intelligent behaviors are likely to emerge.

Instead of further expanding on, or attempting to standardize, the behavioral markers that are believed to be signs of general intelligence, we believe that it may be more fruitful to focus on the necessary characteristics of systems that are likely to produce generally intelligent behaviors. Towards that end, we introduce three of these properties in this paper. In particular, we claim that attempts at engineering generally intelligent systems will likely require the creation of agents that are *autonomous* (capable of acting in pursuit of their own agendas), *cognitive* (capable of reflecting on, and reasoning about their environments in a manner that is decoupled from their immediate inputs and outputs), and *intentional* (capable of connecting the contents of their mental states to corresponding referents in their environments). We contend that deficits in any one of

these properties can lead to pathological behaviors in humans and other animals, and their complete absence is incompatible with general intelligence. To facilitate a discussion on the need for these properties in generally intelligent systems, we provide additional background on what each of these properties entails.

2. Autonomous, Cognitive, and Intentional Agents

The three-factor agent classification presented in this section is similar in spirit to the taxonomy presented by Franklin and Graesser (1997); however, our specific goal is to open a dialog within the community about the necessity of these properties for general intelligence, and the mechanisms by which they can be realized. Towards this goal, we describe each property in detail, along with examples and counter-examples.

2.1 Autonomous Agents

Franklin and Graesser (1997) defined an *autonomous agent* as "a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." To say that an agent has an agenda implies that it is capable of *appraising* environmental states based on its own motivational system, and that it *prefers* some of those states to others. To say that an agent acts "in pursuit of its own agenda," implies that it selects actions *purposefully*, in accordance with that agenda. In other words, autonomous agents must not only have preferences, and the ability to evaluate environmental states with respect to those preferences, they must also have an action selection mechanism that advances the pursuit of those desirable states (though not necessarily in an optimal way).

Examples and Counter-Examples

Reinforcement learning (RL) (Sutton & Barto, 2018) is an agent-based machine learning paradigm in which agents sense their environments, and, though trial-and-error exploration of those environments, learn to choose actions that maximize their "rewards." Rewards are based on an agent's *reward function*, which maps environmental states onto scalar values that quantify that agent's immediate hedonic (liking or disliking) responses to those states. *Model-free* RL algorithms, such as temporal-difference (TD) learning (see Sutton & Barto, 2018), learn to approximate *value functions* that quantify the cumulative (long-term) expected reward associated with each state (or of each action when taken from those states). These value functions make the agent's agenda explicit, and accessible to the agent, and agents that choose their actions based on such value functions are examples of autonomous agents with *explicit agendas*. Other types of model-free RL agents, such as those based on policy-gradient methods (see Sutton & Barto, 2018), do not learn value functions, but instead learn to *directly* optimize their "behavioral policies" (that is, their selection of actions). These RL agents do not use their subjective judgments about the "goodness" of environmental states and actions to guide their action selection, but, instead, simply execute actions that they have learned to be generally useful for satisfying their goals in a given situation. These agents are examples of autonomous agents with *implicit agendas*.

Agents that lack an agenda, or choose actions that are inconsistent with their agendas, are *non-autonomous*. The simplest examples are agents that choose their actions indiscriminately (for

example, randomly). A more subtle example of non-autonomy occurs when an agent's agenda, or action selection mechanism, has been subverted in some way, such as by parasites and diseases (in natural systems), or other source of malfunctions (in artificial systems).

2.2 Cognitive Agents

Franklin and Graesser (1997) defined reactive agents as those that respond immediately to environmental stimuli without consulting an internal model, or engaging in activities such as reasoning, planning, or deliberative thought. By contrast, we refer to agents that are capable of utilizing internal models and thought processes that are detached from current sensory stimuli as *cognitive agents*. To make this distinction more precise, it is useful to consider several concepts from the embodied cognition literature. "Situated cognition" refers to task-specific, context-sensitive, modes of acting that are continually influenced by incoming sensory stimuli. In its most extreme form, situated cognition leads to reactive agents that act through a process of "online" control. "Online" in this sense implies the direct and immediate coupling between an agent's actions and the stimuli that resulted in those actions. "Offline" processes[1], on the other hand, are *decoupled* from an agent's inputs and outputs that are occurring "right now." These processes enable the construction (and manipulation) of imagined realities, as well as forms of "mental teleportation" (that is, spatial decoupling) and "mental time travel" (that is, temporal decoupling). Planning, reasoning, introspection, and problem-solving, as well as more pedestrian activities, like the recall of long-term memories and daydreaming, are all "offline" cognitive processes. Having introduced this terminology, we define a cognitive agent as "an agent that makes sense of, and acts on, its environment based, in part, on processes that are decoupled (temporally, spatially, or otherwise) from its immediate inputs and outputs."

Examples and Counter-Examples

IBM's Watson (Ferrucci, et al., 2010) is a cognitive software agent that defeated the best human contestants of the TV quiz show *Jeopardy* (back in 2011). It was built using over 100 different techniques for "analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses" (Ferrucci, et al., 2010). It analyzed "clues" using "shallow parses, deep parses, logical forms, semantic role labels, coreference, relations, [and] named entities" (Ferrucci, et al., 2010). It generated candidate answers and supporting evidence using multiple text search engines, document and passage-specific search algorithms, and "knowledge base searches." And it applied multiple "scoring algorithms," including those based on geospatial and temporal reasoning, to determine its degree of certainty in those answers.

An example of a non-cognitive agent is a simple thermostat, which reactively turns an air conditioner (or heater) on or off based on its measurements of the current temperature. Brooks's (1990) robots are more sophisticated examples of non-cognitive agents based on the "subsumption architecture." These software agents produce seemingly goal-directed behavior based on the collaboration of many independently operating, reactive "behavior" processes, without centralized control or discernable mental representations.

[1] See Wilson (2002) for an introduction to the idea of offline cognition, and its relationship to embodied cognition.

2.3 Intentional Agents

A fundamental problem that confronts system designers is how to establish a correspondence between internal mental representations and what they represent in the external world. We refer to this as the problem of *intentionality*[2] (that is, "aboutness") (Searle, 1980); however, this issue has more commonly been referred to as the *symbol grounding problem* (Harnad, 1990). The symbol grounding problem is best illustrated by considering classical symbolic AI systems, which were based primarily on the explicit, rule-based manipulation of *symbolic representations*. These representations are *by definition*[3] arbitrary, as their forms do not depict or resemble their referents, and they bear no intrinsic informational content that would suggest a connection between them and the concepts to which they refer. Harnad (1990) illustrated the symbol grounding problem by offering, as an example, the formidable task of trying to learn Chinese as a first language when the only information at your disposal is a Chinese-to-Chinese dictionary: "[using] the dictionary would amount to a merry-go-round, passing endlessly from one meaningless symbol… to another… never coming to a halt on what anything meant" (Harnad, 1990). While we may learn an elaborate web of correlations between different words, we will never establish the meaning of those words based solely on their associations with other unintelligible words.

The symbol grounding problem is often glossed over in practice because humans are generally "in the loop" to interpret (that is, give meaning to) the results of a machine's computational efforts. This effectively connects symbols with their meanings *exogenously*, and after the fact. If all we care about is creating useful software tools, such as a spelling and grammar checker, then the fact that the system does not understand what those text strings mean is irrelevant. However, we contend that any general intelligence must be able to establish a correspondence between its mental representations and their underlying concepts in the world. We call agents that can determine such connections between internal mental representations and the external world *intentional agents*. Note that since intentionality refers to the link between an agent's representational mental states and their referents in an environment, agents must have representations to be considered intentional.

Examples and Counter-Examples

Any agent that is implemented using a *purely symbolic approach* is non-intentional, and requires a human (or other external entity) to connect its symbols to the concepts they signify. Perhaps as a result of this issue, and the need for systems to operate in more complex environments, a majority of cognitive architectures have adopted a hybrid symbolic/non-symbolic approach (Kotseruba & Tsotsos, 2018). What may be less obvious is that many software systems based on non-symbolic (for example, connectionist) approaches are also non-intentional.

BERT (Devlin, Chang, Lee, & Toutanova, 2018), which stands for ***B**idirectional **E**ncoder **R**epresentations from **T**ransformers*, is an artificial neural network (ANN) architecture that achieved state-of-the-art performance (circa 2018) on many "natural language understanding" tasks. BERT's inputs consist of sentences, or pairs of sentences, where "sentence," in this context,

[2] Jacob (2020) defined intentionality as "the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs." It is important to note that this use of the term intentionality is different than its colloquial use to denote deliberate or purposeful activities.

[3] This terminology was established by Peirce in the late 19th century as part of his theory on semiotics (Houser & Kloesel, 1992).

refers to an "arbitrary span of contiguous text." BERT was pre-trained on over 3 billion words, and the resulting model was fine-tuned to solve 11 different natural language tasks. Some of these tasks required judgments about the semantic equivalence and similarity of sentences, others about author "sentiment" or the correctness of grammar, and still others required choosing the most plausible "continuation sentence" for a given example sentence. On some of these tasks BERT *outperformed* human experts. Can we conclude that BERT knows what those sentences "mean"? *Absolutely not.* After being trained on billions of words, the network has learned linguistic regularities, such as word correlations and morphology, but it does not understand what those words signify in the world. That is not to say BERT is not a useful tool for humans. Far from it! But it is not an intentional system capable of endogenous meaning. To be an intentional system, BERT would also need to incorporate non-symbolic inputs (that is, worldly experiences such as images, sounds, etc.) corresponding to each (or at least some) of its symbols.

3. General Intelligence Requires Autonomous, Cognitive, Intentional Agents

Having established our working definitions of *autonomous*, *cognitive*, and *intentional agents*, we now claim that general intelligence likely requires that agents have *some degree* of all three properties.

The Case for Autonomous Agents. We regard the joint concepts of autonomy and agency, as reflected in Franklin and Graesser's (1997) definition of autonomous agent (presented in Section 2.1), as a minimal starting point for intelligence, and contend that minds are best defined as "control structures for autonomous agents" (Franklin, 1997). We believe that humans, and other animals, with reduced autonomy may suffer from *anhedonia* (inability to feel pleasure, or a loss of interest in engaging in activities), *attentional disorders* (inability to focus on task-relevant stimuli, or to complete activities), or *motor disorders* (such as Tourette's syndrome that results in unwanted and involuntary actions). Based on this, we contend that the complete absence of autonomy is surely inconsistent with general intelligence.

The Case for Cognitive Agents. Most definitions of general intelligence contain an explicit or implicit behavioral requirement that the system be capable of performing a variety of demanding tasks in complex environments, and developing innovative, system-appropriate, solutions to the needs of those environments. Since non-cognitive agents (see Section 2.2) are incapable of projecting their thoughts beyond the immediate present, we argue that it is unlikely that attempts to scale up such reactive agents to solve a variety of complex problems in open-ended, complex environments will be successful. We believe that humans, and other animals, with reduced cognitive abilities are often characterized as being *short-sighted* (fixated on the immediate present), *impulsive* (reckless and unaware of consequences), *irrational* (unable to "consciously" select actions consistent with their desires), or *unimaginative* (unable to synthesize new ideas). Based on this, we contend that the complete absence of "offline" cognitive abilities is surely inconsistent with general intelligence.

The Case for Intentional Agents. We regard an agent's ability to connect its internal mental representations to environmental referents as a fundamental requirement for differentiating software tools, which are oblivious to the significance of their labors, from more self-aware and comprehending machines. If an agent's mental representations have no intrinsic meaning, then they will always be dependent on exogenous entities (such as humans) to give meaning to their outputs. We believe that humans, and other animals, with reduced intentionality may be considered *disoriented* (suffering from a loss of time, place or identity), *delusional* (maintaining

false beliefs or perceptions despite clear evidence to the contrary), *psychotic* (disconnected from reality and unable to distinguish real from unreal), or suffering from *agnosia* (inability to recognize objects, places, or situations from sensory stimuli). Based on this, we contend that the complete absence of intentionality is surely inconsistent with general intelligence.

4. References

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and autonomous systems, 6*(1-2), 3-15.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI magazine, 31*(3), 59-79.

Franklin, S. (1997). *Artificial Minds.* MIT Press.

Franklin, S., & Graesser, A. (1997). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages* (pp. 21-35). Springer-Verlag.

Goertzel, B., & Pennachin, C. (2007). The Novamente artificial intelligence engine. In B. Goertzel, & C. Pennachin (Eds.), *Artificial general intelligence* (pp. 63-129). Springer, Berlin, Heidelberg.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*, 335-346.

Houser, N., & Kloesel, C. (Eds.). (1992). *The essential Peirce, Volume 1: Selected philosophical writings (1867–1893).* Indiana University Press.

Jacob, P. (2020). Intentionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2019/entries/intentionality/.*

Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1-78.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communication of the ACM, 19*(3), 113-126.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences, 3*(3), 417-424.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Voss, P. (2007). Essentials of general intelligence: The direct path to artificial general intelligence. In B. Goertzel, & C. Pennachin (Eds.), *Artificial general intelligence* (pp. 131-157). Springer, Berlin, Heidelberg.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review, 9*(4), 625-636.