There Is No Agency Without Attention

Paul F. Bello, Will Bridewell

■ *Over the decades, the view of agency in artificial intelligence (AI) has narrowed to one that emphasizes acting in a way that maximizes reward. This perspective fails to make contact with the broader academic and legal communities where agency is bound up with personal accountability. To explore this gap in meaning, we introduce a spectrum of control that characterizes standard approaches to constructing agents and points the way toward agents that can be held responsible. The linchpin that enables agents to control their actions in the "right way" is attention. Broadly construed, attention lets an agent that is responsive to its environment consider the relationships among its actions, goals, and norms while also avoiding distraction. This ability enables strategic norm violations and opens the door to artificial, human-level agency.*

Herb Simon (1996) begins the third chapter of *The Sciences of the Artificial* with a parable about an ant making her way home across a beach and comments that the complexity of her route is not a manifestation of any complex, goal-seeking behavior within the ant but rather reflects the complexity of the environment. The ant "must adapt [her] course repeatedly to the difficulties [she] encounters and often detour uncrossable barriers." The question that interests us is not whether the ant is complex, but whether she should be considered an intelligent agent. The ant appears to have a goal to get home, and she exhibits control in her ability to navigate novel and complex terrain. Her control is grounded in a coupling of her perceptions to her actions that keep her from continually bumping into an obstacle. Nevertheless, calling the ant an intelligent agent seems like overzealous anthropomorphism on our part. Why is that?

Within the artificial intelligence (AI) community, this question is not new. For decades AI researchers have built agents that are capable of carrying out tasks that require human-level or humanlike intelligence. During this time, questions of how these programs compared in kind to humans have surfaced and led to beneficial interdisciplinary

 ISSN 0738-4602

discussions, but conceptual progress has been slower than technological progress. Over the past decade, the term *agency* has taken on new import as intelligent agents have become a noticeable part of our everyday lives. Research on autonomous vehicles and personal assistants has expanded into private industry with new and increasingly capable products surfacing as a matter of routine. This wider use of AI technologies has raised questions about legal and moral agency at the highest levels of government (National Science and Technology Council 2016) and drawn the interest of other academic disciplines and the general public. Within this context, the notion of an intelligent agent in AI is too coarse and in need of refinement. We suggest that the space of AI agents can be subdivided into classes, where each class is defined by an associated degree of control.

On this front, we can identify three categories of agents based on differences in how control is manifested given a number of factors. To a first approximation these factors include the agent's ability to systematically cope with ignorance or indifference by generating and evaluating hypotheticals and the agent's ability (1) to choose a course of action against the background of explicit norms for action selection and (2) to remain committed to that course in the presence of ordinary distractions (Cohen and Levesque 1990). In general, the three different classes of agents trend toward greater flexibility by being less bound to peculiarities of the current situation and less constrained by whichever principles determine action on a moment-to-moment basis. As we give examples below, we will do our best to point out differences in these dimensions.

Three Types of Agents

Agents of the first type have fixed representations of their environment, including environmental uncertainty. By this, we mean that an agent knows about all the variables, objects, and events in its environment and can effectively build a state space from this information even if initial values for the variables are unknown. Actions for this kind of agent are essentially stimulus-response mappings that may be determined by hard-wired principles such as utility maximization, cost minimization, winner-take-all, or through a fixed ordering scheme. In this case, control involves keeping a known set of state variables within legal ranges or responding to changes in state variables in specific ways. Canonical examples include a home's thermostat, a car's cruise control, and an airplane's autopilot system.

Agents of the second type are flexible in the face of ignorance or indifference. Rather than being able to represent only what they know to be the case given background knowledge, they can generate what might be the case hypothetically and can respond to evaluations of those hypotheticals. Consider a deer when thirst drives her to drink from a stream and the smell of a bear drives her to run to the forest. Whether she drinks or flees depends on interactions among several internal and external factors, including her calculated distance to refuge (Stankowich and Blumstein 2005). Importantly, there is flexibility based on which of those factors draw her attention (Bernays and Wcislo 1994). While the action-determining principles used by the second type of agent may not differ in kind from those used in the first type, the generation of hypotheticals dynamically expands the state space and enables a measure of control otherwise unattainable.

Agents of the third type are capable of committing to choices with respect to norms. The defining characteristic of norms is that they specify what an agent ought to do rather than determining what an agent actually does. Explicitly represented norms can subsume the sort of fixed principles that determine the actions taken by the agents of the first and second types. So while those two types of agents unfailingly act in ways that maximize utility or minimize cost, agents of the third type have these principles encoded explicitly. This explicit encoding makes the norms available for reasoning, comparison, and interchange based on the dynamics of the situation. Considered as a norm rather than hard-wired mechanisms, cost minimization might be violated in light of circumstances in which the expected costs of performing various actions are unknown whereas the expected utilities of potential outcomes are well characterized. Similarly, there may be situations in which the utilities of two different actions are incommensurate, such as when one of the actions has moral significance and the other is driven by self interest. For instance, a hungry man who forgot his wallet may be inclined to take a sandwich from a shop counter. In this situation, he may satisfy his hunger without much thought, he may recall his morals that prohibit stealing and opt to walk away, or he may choose to violate or reinterpret his moral code and steal the sandwich.

Along with their actions being guided by explicit norms, agents of the third type commit themselves to pursuing their goals in the face of distractions, situational reappraisals, and so on. Such an agent might have any number of opportunities to abandon its longer-terms plans in favor of a shorter-term opportunity for gain, or it may find goal pursuit stymied by distracting stimuli.[1] A committed agent must have mechanisms that allow it to remain focused on goal pursuit so as to avoid vacillation. We claim that these focusing mechanisms and the ability to use norms in action selection are both necessary conditions for human-level agency.

What kind of mechanisms could plausibly account for differences between all three types of agent? Our answers are attention and control of attention. Before we discuss attention and its role in control, we explore and contrast various definitions of agency

and control that led us to propose the particular taxonomy sketched in this introduction. As we progress, we find that a rich capacity for agency imposes certain demands on the design of cognitive systems. By the end of our discussion, we anticipate that the reader will agree with our claim that *there is no agency without attention,* and that a rich account of attention ought to be at the heart of any cognitive system that aims to exhibit human-level agency.

What Is Agency?

In the first AAAI Presidential Address, Allen Newell spoke at length about intelligent agents. In the published version of his lecture he wrote,

> … an agent is composed of a set of actions, a set of goals and a body … the agent processes its knowledge to determine the actions to take … the behavior law is the principle of rationality. Actions are selected to attain the agent's goals (Newell 1981, p. 6).

With an emphasis on the interactions among knowledge, goals, and actions, Newell's knowledge-level perspective continues to inspire researchers in cognitive systems. For comparison, the last 20 years has seen the broader AI community treating a rational agent as one that "for each possible percept sequence … should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has" (Russell and Norvig 2002, p. 36). This emphasis on maximization conflicts with interdisciplinary and nontechnical views on agency in a way that Newell's treatment does not, but neither perspective captures the richness of agency as it is broadly understood.

Consider first the definition from Russell and Norvig. For them, a rational agent should be built to act unfailingly in the right way, given all its knowledge. This implies that agents of this sort can act only on the information they have, which may include representations of known uncertainty. Maximization or minimization principles require complete characterization of the state of the world. However, there are many cases where action must be taken in the absence of information — when uncertainty cannot be quantified or when an agent's environmental model fails. In such cases, clear preferences among alternatives may never have been learned or may not be possible to uncover.

Other situations raise concerns for both Russell and Norvig's and Newell's accounts of agency. Suppose that a person gets up, puts on their clothes, drives to their in-laws' house, grabs a knife from the kitchen, stabs their mother-in-law to death, and then drives to the police station claiming to have killed someone. On this description, we would say the person is not only an intelligent agent but also a moral agent, one who can bear responsibility for their actions (see Scheutz [2017]).[2] Now, imagine that same scenario but where the person involved was somnambulant, carrying out the actions while asleep. This suggestion may sound like a philosopher's game, but somnambulant assaults and homicides are well documented. The aforementioned sketch roughly corresponds to the case of Kenneth Parks (Broughton et al. 1994) whose acquittal from homicide was upheld under appeal. The legal argument was that Parks exhibited noninsane automatism, which requires the offending action(s) to be unconscious and involuntary.[3] Parks's ability to drive a car, search for a knife, and attack his in-laws suggests that he was an intelligent agent, but something was missing from his experience that prevented him from being a moral agent.

What do the accounts of agency presented above have to say regarding this example? The answer, unfortunately, is, "not much." Neither Russell and Norvig's nor Newell's definition can distinguish between the normal and somnambulent versions of the homicide story. Agents as defined by both views lack the ability to "act otherwise" — their actions are inherently unconscious and involuntary. In contrast, agents who choose to act do so (typically) voluntarily and consciously. If the sleepwalker, who is unaware of his actions and unable to veto them, is not accountable, then neither is any form of AI agent that fails to distinguish between or provide some analogue to conscious and unconscious content. Similarly, an agent cannot be held accountable if it lacks reliable mechanisms capable of bringing its behavior in line with its consciously considered desires, obligations, and the like.

Does this mean that the development of artificial, accountable, moral agents necessitates a computational analogue to human awareness? If so, then we should not be hopeful for a ready solution. Although our personal awareness seems immediate and clear, its root causes are nebulous and its contents are barely understood. But we are optimists. Perhaps there is an informational component to awareness that sufficiently supports accountability; a component that a computational model could provide. We can then ask the question that is central to this article. Is there a way to develop intelligent, artificial agents that are also moral agents — agents that have sufficient autonomy to be held accountable? The answer, we claim, is, "Yes," and lies in operationalizing control.

What Is Control?

In the introduction, we identified three kinds of agent-control pairings that we loosely associated with thermostats, deer, and people. In this section, we call these $control_0$, $control_1$, and $control_2$, respectively. Essentially all AI systems that act in the world exhibit $control_0$. The requirements at this level include little more than a learned policy or a set of production rules. Typically, policies like these are

learned with respect to hardwired principles such as utility maximization or cost minimization. For the purpose of contrasting these principles with our norm-centered account of *control*$_2$, we refer to them as implicit control knowledge. If simultaneously occurring, mutually exclusive, action requests are even possible in these systems, the conflicts are resolved through inflexible techniques like static rule-ordering or a coin flip. In essence, such systems are effectively a collection of stimulus-response mappings and the only basis upon which to call them "agents" is in virtue of the fact that they act. Moreover, to say that such agents "select" actions is incorrect. Stimulus-response mappings generated by implicit control knowledge determine what actions such systems take.

Control$_1$

Fewer AI systems implement techniques to support *control*$_1$. This level of control requires flexibility under conditions of ignorance or indifference. Flexibility in this sense is a matter of degree, but to illustrate the idea, we look at how a particular cognitive system addresses this requirement. Soar is a cognitive architecture for developing agents (Laird 2012). At its heart, the architecture is a production system that operates in cycles. At the risk of oversimplification, during each cycle a Soar agent acquires percepts, elaborates these through its encoded knowledge, and uses the subsequent state to select an operator for application. Importantly, a Soar agent can apply only one operator per cycle, and these operators are what initiate action in the world. Occasionally, Soar's (*control*$_0$) preference mechanisms cannot select among multiple operators on a cycle. In this case the architecture creates a substate in which it can carry out forward simulation while continuing to monitor the environment. By using this mechanism to identify a preferred course of action, the agent resolves its conflicts and exhibits *control*$_1$. Other AI systems may use different mechanisms, but the important feature is the ability to acquire or construct new, situation-specific information that will influence the selection of one action over another.

Control$_2$

Control$_2$ expands the flexibility of *control*$_1$ by (1) making control knowledge explicit and (2) providing a means for agents to remain committed to goal pursuit in the face of distractions, temptations, and situational reappraisals.

Norms as Explicit Control Knowledge

Over the course of the discussion, we have emphasized the notion that maximization or minimization principles used for agents exhibiting *control*$_0$ are norms or that they at least express something like normative content. This aside, agents exhibiting *control*$_0$ need never reason directly about such norms and their applicability. These built-in norms are implicit and determine how an agent acts rather than being explicit and defeasible suggestions as to how the agent ought to act. Expanding on our earlier description, norms are general principles or schemata coupled with procedures for elaboration and interpretation that make them applicable to specific situations. To guide behavior, norms must be recalled from memory, instantiated to the situation, assessed for applicability, and probed for exceptions, perhaps at times through analogy (Forbus and Hinrichs 2017). Because multiple conflicting norms may apply (for example, do not steal and do not starve), an agent may need to invoke strategies for resolving norm conflicts. These strategies may weigh norms based on their provenance (for example, religious injunctions may take precedence over workplace protocols) or may suggest modifications to actions that lead to acceptable behavior. As a loose example of norm-based reasoning, consider how the hungry man might align his actions with his norms:

> Taking a sandwich is stealing because there is no permission for the act.
> Buying a sandwich confers permission to take it.
> Other ways to gain permission are to ask or negotiate.
> Therefore, negotiating to pay for the sandwich later could enable taking the sandwich without stealing.

This case does not involve the *control*$_1$ activity of comparing competing actions or goals (for example, Johnson et al. [2017]). Instead, this example illustrates how norms guide an agent to reflect on available actions or goals and to explore modifications that produce norm-conforming behavior. Exhibiting *control*$_2$ does not always require this sort of creative effort, but it does require the ability to apply general standards of behavior to specific situations.

Commitment

The case of Kenneth Parks provides reasonable grounds to distinguish between actions consciously chosen by an agent and activity merely determined by unconscious content. But an agent's choice is only as good as its commitment to follow through. Our commitments to actions or goals, such as eating an apple instead of a donut or maintaining a healthy diet, can be fragile especially in the face of temptations, distractions, and situational reappraisal. Our perceptions influence our actions and can take us off course. This relationship between perception and action is built into both Newell's knowledge-level agents and Russell and Norvig's rational agents. Fortunately, we can influence our perceptions by controlling our actions. If a person turns her head, her visual field changes. If she dons headphones, she alters her soundscape. If she is committed to eating healthily, she may move donuts out of sight or avoid places where they are served. This is achievable when an agent (1) has explicit causal knowledge about how perception, mental states (for example, beliefs, obligations, desires), and action are related and (2) can

use this information to predict which actions might reduce distractions or temptations. Some of these actions might involve changing the nature of the agent-environment interaction, such as closing a window to prevent street-level noise from interrupting writing a paper. Other actions may be mental in nature, like avoiding a distracting negative thought by mentally rehearsing a song or the weekly shopping list.

Taken together, the ability to choose a course of action and to commit to its execution enables the intentional behavior that some believe is necessary for autonomous agents (Cohen and Levesque 1990). For us, what is important is that when an agent is instructed to carry out a task, it can choose to do otherwise, as that is the ability that enables human-level agency. We claim that a single mechanism that exists to some degree in $control_1$ agents is central to the capacities of choice and commitment in $control_2$ agents. That mechanism is attention.

Attention and Control in Cognitive Systems

To reiterate, we have argued that intelligent agency is intimately related to control and that a special kind of control is required for human-level agency. We also made a case that this form of control requires both a mechanism of attention and knowledge that can predict the mechanism's effects. But what does attention do in a cognitive system such that it might enable control? In prior work, we have argued that six particular desiderata, when jointly fulfilled by a cognitive system, are sufficient for claiming that the system has a rich capacity for attention, which (1) is limiting and selective; (2) can be directed inward or outward; (3) can be captured or intentionally directed; (4) asymmetrically biases mental processing; (5) facilitates integrated mental processing; and (6) facilitates conscious access. Our earlier report justifies these features (Bridewell and Bello 2016), and we take them at face value throughout the rest of our discussion.

To what degree do the three types of agents fulfill these desiderata? A $control_0$ agent is designed to not need attention. These agents operate over a fixed set of actions and state variables that are preselected and rely on static preferences that determine which actions take priority at any time. In contrast, a $control_1$ agent has a mechanism of attention that enables flexible goal pursuit. As we saw in the previous section, a Soar agent can resolve impasses through forward simulation, which shifts its attention to conflicting operators and away from the routine application of actions. When this happens, attention mediates the link between perception and action execution in an agent, enabling it to compare alternative actions based on their expected effects and their relationship to that agent's goal. In terms of the desiderata, the mechanisms of impasse resolution and hypothetical evaluation satisfy items 1–3. Specifically, Soar (1) selectively elaborates operators (2) in substates (3) in the case of impasses. This process intervenes to break the perception-action link and enables control in the face of ignorance or indifference.

Although the abilities to interrupt the perception-action cycle and to consider alternatives are necessary aspects of attention, they are insufficient for enabling $control_2$. As it stands in Soar, nothing about the substate reasoning process distinguishes rote actions from actions open to reflection by an agent. More directly, even on a thin, computational notion of awareness that maps to levels of information availability, Soar agents lack awareness of any information including any norms that guide their behavior. Notably, such a minimal form of awareness is sufficient to distinguish the two cases of homicide presented earlier, where opening both actions and norms to deliberation would have let Kenneth Parks evaluate and veto his aberrant behavior. So, even though Soar agents satisfy items 1–3 in the desiderata, they lack item 6, which is necessary to support conscious choice.

Attention and Norms

We submit that no artificial agent can properly choose unless alternatives are represented explicitly in some analog to consciousness. Further, we contend that explicit control knowledge given in the form of norms is also represented in this analog. To be clear, we do not disavow the possibility or the potential efficacy of implicit, unconscious norms in guiding behavior; such ideas stretch as far back as Aristotle and form the foundation for entire classes of ethical theories. Rather, we emphasize explicit norms because they can be invoked as reasons, which feature prominently in theories linking freedom of the will, agency, control, and responsibility (Fischer and Ravizza 1998). Furthermore, psychological data suggests that norms are central to negotiating blame and are thus a crucial ingredient for determining accountability (Malle, Guglielmo, and Monroe 2014).

What is the relationship between attention and the application of norms? Attention must be directed toward both the interpretation of the norm and the hypothetical consideration of an action within a situation. At the surface level, norms and actions are not necessarily comparable. Specifically, to compare the base action of taking some item to a norm against stealing, an agent must answer a variety of questions.

1. Does the norm apply to items of this type?
2. Does the norm apply to this specific item?
3. Is there explicit permission to take the item?
4. Is there implied permission to take the item?
5. What are the expected results of the action?

This list, which is not comprehensive, includes top-down questions that elaborate the norm within

a situation (1 and 2) and "bottom-up" questions that position the action within a larger system of goal-driven and norm-guided behavior (3–5). Notably, a question may require further direction of attention. For instance, determining whether permission is implied requires guided reasoning, which is impossible without sustained attention.

Attention and Commitment

Previously we noted that an agent's choice is only as good as its commitment to follow through. One of the primary threats to commitment is distraction, which presents a challenge to $control_1$ agents whose attention is directed only accidentally. These agents rely on a fixed procedure for carrying out inference and action selection, and that procedure's input cannot influence its operation in any way. This limitation poses a problem for committing to actions and goals because there are no means to avoid perceptions or actions that would lead to broken commitments. Suppose someone committed to eating healthily decides to stop into the nearest cafe, which also sells donuts. While there, she notices that she is hungry and buys a donut because it is the only available action to satisfy her hunger. If she had reflected on her commitment she may have avoided the donut shop altogether or possibly adopted an explicit goal to not buy a donut, which would guard against an impulse purchase. Knowing how particular actions will influence perception and knowing how to manipulate attention (for example, by adopting a goal and maintaining it) enables a $control_2$ agent to successfully and durably commit to its choices.

Our task was to argue that attention is necessary for control and to point out the details of that connection, but we are not the first to argue this point. In the psychological and neuroscientific literature, we find like minds in the research that investigates the connection between attention and action (Allport 1987; Hommel 2010; Wu 2011). Moreover, in AI the importance of control knowledge, or "strategies to guide the use of knowledge" (Davis and Buchanan 1984), was well appreciated when there was greater emphasis on what are now called cognitive systems. Moving forward, emphasizing attention's necessary place in any account of agency will increase research interests in this area and will ensure it a place of prominence in any unified theory of the mind (Laird, Lebiere, and Rosenbloom 2017).

Future Directions

We have gone some distance toward an account of agency suitable to be explored computationally within cognitive systems. Specifically, we have provided a set of desiderata relating to attentional capacities, along with some attendant representational assumptions that are required for the exertion of control. Our account of control relies on three key elements: (1) information parcelled into sets that distinguish potentially conscious from unconscious content, (2) an attentional mechanism that influences whether content is consciously available, and (3) explicit control knowledge that can influence the attentional mechanism.

However, many questions remain unanswered. As we have developed an account of control, we have presumed that something was doing the controlling. Agency presupposes an agent, and we have not given any account of the first-personal qualities of agency. When a particular student exercises agency by raising her hand in class, what explains how she knows it is her arm that is raised and that she did the raising, especially when she is surrounded by other students raising their hands? In general, we have not exhaustively specified the conditions under which the agent mentally represents itself as acting versus situations where the agent acts in the absence of self-representation. We have been silent about just how actions under consideration might be made consciously available, noting only the role of attention in this process. Also missing is a description of how considered actions turn into intentions, which lead to action execution. Each of these topics deserves a full treatment that we cannot provide here. However, we are pursuing these concerns in our own research on ARCADIA (Bridewell and Bello 2016) — a system that provides functionality that meets the six desiderata for attention and the three elements critical for $control_2$.

Concluding Remarks

In summary, a $control_2$ agent is capable of choice because it can attend to the relationship among actions, goals, and norms and is capable of commitment because it can direct attention to guard against sources of distraction and temptation. The ability to interpret the norms that guide behavior enables a degree of choice not found in $control_0$ or $control_1$ agents. An agent may, depending on its immediate assessment of a norm and available actions, choose to act in a way that violates the norm. Taken together, the ability to choose a course of action and to commit to it enables the intentional behavior that some believe is necessary for autonomous agents (Cohen and Levesque 1990). For us, what is important is that when an agent is instructed to carry out a task, it can choose to do otherwise, as that is the ability that enables human-level agency. With this position in mind, we can clarify our initial claim and state, "there is no human-level agency without control of attention."

Acknowledgments

We thank Richard Weyhrauch for early conversations that influenced the ideas on control and attention. We also thank Bruce Buchanan, Alistair Isaac, John

Laird, Pat Langley, and Sergei Nirenburg for discussion and comments that led to improvements in the text. The authors would like to acknowledge support from the Office of Naval Research under grants N0001416WX01112, N0001417WX00153 and N000 1416WX00762. The views expressed in this article are solely those of the authors and should not be taken to reflect any official policy or position of the United States government or the Department of Defense.

Notes

1. We assume a broad view of sensing here that allows for thoughts, memories, and other subjective features that might be distracting.

2. It should go without saying that human agency, at least typically, includes the capacity for moral agency.

3. The Supreme Court of Canada wrote in its judgment, "It may be that some will regard the exoneration of an accused through a defence of somnambulism as an impairment of the credibility of our justice system. Those who hold this view would also reject insane automatism as an excuse from criminal responsibility. However, these views are contrary to certain fundamental precepts of our criminal law: *only those who act voluntarily with the requisite intent should be punished by criminal sanction"* (italics added, quoted in Broughton et al. 1994).

References

Allport, D. A. 1987. Selection for Action: Some Behavioral and Neurophysiological Considerations of Attention and Action. In *Perspectives on Perception and Action*, ed. H. Heuer and A. F. Sanders, 395–419. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bernays, E. A., and Wcislo, W. T. 1994. Sensory Capabilities, Information Processing, and Resource Specialization. *The Quarterly Review of Biology* 69(2): 187–204. doi.org/10.1086/418539

Bridewell, W., and Bello, P. 2016. A Theory of Attention for Cognitive Systems. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems,* 1–16. Palo Alto, CA: Cognitive Systems Foundation.

Broughton, R.; Billings, R.; Cartwright, R.; Doucette, D.; Edmeads, J.; Edwardh, M.; Ervin, F.; Orchard, B.; Hill, R.; and Turrell, G. 1994. Homicidal Somnambulism: A Case Report. *Sleep* 17(3): 253–64.

Cohen, P. R., and Levesque, H. J. 1990. Intention Is Choice with Commitment. *Artificial Intelligence* 42(2–3): 213–261. doi.org/10.1016/0004-3702(90)90055-5

Davis, R., and Buchanan, B. G. 1984. Meta-Level Knowledge. In *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project,* ed. B. G. Buchanan and E. H. Shortliffe, 507–530. Reading, MA: Addison-Wesley Publishing Company.

Fischer, J. M., and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9780511814594

Forbus, K., and Hinrichs, T. 2017. Analogy and Relational Representations in the Companion Cognitive Architecture. *AI Magazine* 38(4). doi.org/10.1609/aimag.v27i2.1882

Hommel, B. 2010. Grounding Attention in Action Control: The Intentional Control of Selection. In *Effortless Attention: A New Perspective in the Cognitive Science of Attention and Action,* ed. B. Bruya, 121–140. Cambridge, MA: The MIT Press. doi.org/10.7551/mitpress/9780262013840.003.0006

Johnson, B.; Coman, A.; Floyd, M. W.; and Aha, D. W. 2017. Goal Reasoning and Trusted Autonomy. In *Foundations of Trusted Autonomy,* ed. H. Abbass, J. Scholz, and D. Reid. Berlin: Springer.

Laird, J.; Lebiere, C.; and Rosenbloom, P. 2017. A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2744

Laird, J. E. 2012. *The Soar Cognitive Architecture.* Cambridge, MA: The MIT Press.

Malle, B. F.; Guglielmo, S.; and Monroe, A. E. 2014. A Theory of Blame. *Psychological Inquiry* 25(2): 147–186. doi.org/10.1080/1047840X.2014.877340

National Science and Technology Council. 2016. Preparing for the Future of Artificial Intelligence. October 2016. Washington, D.C.: Executive Office of the President, Committeee on Technology. (www.whitehouse.gov/sites/default/files/whitehousefiles/microsites/ostp/NSTC/preparing for the future of ai.pdf.) Accessed October 12, 2016.

Newell, A. 1981. The Knowledge Level: Presidential Address. *AI Magazine* 2(1): 1–20. doi.org/10.1609/aimag.v2i2.99

Russell, S. J., and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach,* 2nd edition. Upper Saddle River, NJ: Prentice Hall.

Scheutz, M. 2017. The Case for Explicit Ethical Agents. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2746

Simon, H. A. 1996. *The Sciences of the Artificial,* 3rd edition. Cambridge, MA: The MIT Press.

Stankowich, T., and Blumstein, D. T. 2005. Fear in Animals: A Meta-Analysis and Review of Risk Assessment. *Proceedings of the Royal Society B: Biological Sciences* 272(1558): 2627–2634. doi.org/10.1098/rspb.2005.3251

Wu, W. 2011. Confronting Many-Many Problems: Attention and Agentive Control. *Noûs* 45(1): 50–76. doi.org/10.1111/j.1468-0068.2010.00804.x

Paul F. Bello is the director of the Interactive Systems Section at the U.S. Naval Research Laboratory, and the former director of the Cognitive Science program at the Office of Naval Research. His research interests lie at the interface between attention, perception, reasoning, and action with a particular focus on consciousness and moral agency. He received his Ph.D. in cognitive science, M.S. in computer science, and B.S. in both computer engineering and philosophy from Rensselaer Polytechnic Institute. He is the codesigner of the ARCADIA attention-driven cognitive system and codirects the ARCADIA research program.

Will Bridewell is a computer scientist at the U.S. Naval Research Laboratory. Formerly he was a research scientist at Stanford University. His current research investigates the relationship between attention and intentional action with a broader interest in computational theories of consciousness. He holds Ph.D. and M.S. degrees in computer science from University of Pittsburgh and B.S. degrees in psychology, mathematics, and computer science from Northern Kentucky University. He is the codesigner of the ARCADIA cognitive system and codirects the ARCADIA research program.