Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 1 (2020) 15–24

DOI: 10.1142/S2705078520500022

World Scientific
www.worldscientific.com

Attention and Consciousness in Intentional Action: Steps Toward Rich Artificial Agency

Paul Bello* and Will Bridewell†

Code 5512, Naval Research Laboratory
4555 Overlook Ave. S.W.
Washington, District of Columbia 20375, USA
**paul.bello@nrl.navy.mil*
†*will.bridewell@nrl.navy.mil*

Published 29 April 2020

If artificial agents are to be created such that they occupy space in our social and cultural milieu, then we should expect them to be targets of folk psychological explanation. That is to say, their behavior ought to be explicable in terms of beliefs, desires, obligations, and especially intentions. Herein, we focus on the concept of intentional action, and especially its relationship to consciousness. After outlining some lessons learned from philosophy and psychology that give insight into the structure of intentional action, we find that attention plays a critical role in agency, and indeed, in the production of intentional action. We argue that the insights offered by the literature on agency and intentional action motivate a particular kind of computational cognitive architecture, and one that hasn't been well-explicated or computationally fleshed out among the community of AI researchers and computational cognitive scientists who work on cognitive systems. To give a sense of what such a system might look like, we present the ARCADIA attention-driven cognitive system as first steps toward an architecture to support the type of agency that rich human–machine interaction will undoubtedly demand.

Keywords: Attention; Agency.

1. Introduction

After an argument with his nephew Eddie, George retires to his bedroom and slips under the covers with Linda, his new beau. Eddie, who has been taking care of George in his old age, cautioned George as he noticed all of the expensive things that Linda has been purchasing with George's money. Looking over at Linda, George becomes even angry with Eddie for thinking Linda as a conniving gold-digger. At that moment, George decides to give Eddie's inheritance to Linda. A week later, a letter arrives for Eddie announcing that he has been excised from George's will and that his services would no longer be required. Eddie flies into a rage and decides to drive to George's house and kill them both. He grabs his shotgun, hops in his car, and begins to drive. On his way over, Eddie is replaying his last heated conversation with his

uncle about Linda, and his receipt of the news that he had been cut out of his uncle's will after all of the time and energy he had spent on caring for old George. In no time, Eddie gets into a feverish rage. Soon his mind was preoccupied with bloody scenes and plans to clean up the mess, but then out of nowhere the car slams into something. Eddie is broken from his dark fantasizing to find that he has run over a pedestrian. He exits the car to investigate and to his surprise, the person he has run down is George! Did Eddie kill George intentionally?

We turn now to a second case study: that of Kenneth Parks. One evening, a still-sleeping Mr. Parks arose from the couch, got dressed, got in the car, drove 14 miles and navigated three different traffic lights to arrive at his in-laws' home. He walked into the house, strangled his father-in-law into unconsciousness, and stabbed his mother-in-law multiple times in the chest and shoulder. She later died of her wounds. He left his in-laws' house and drove to the police station, where he told the authorities that he thought that he had hurt someone. He was horrified to find out what he had done. His memories were fragmentary, with only images of watching TV on his couch, his mother-in-law's horrified face, the knife in his hand, and a few other highly emotional bits and pieces, but nothing connecting the pieces together. Parks was initially charged with first-degree murder, but pleaded not guilty, citing "non-insane automatism," which in layman's terms is effectively sleepwalking. Parks had a history of disturbed sleep, both walking and talking in his sleep, as well as sitting up in bed with his eyes wide open yet being totally unresponsive. Parks was later assessed as displaying parasomnic tendencies, and the various stresses he had been under served as a good explanation of why this particular episode occurred. The fragmentary structure of his recollections of the incident was also consistent with sleepwalking as an explanation. After an extensive investigation, Parks was acquitted of murder, and the acquittal was upheld by the Supreme Court of Canada. The lynchpin to the judgment lies in the dependency between "being aware of what you're doing" and acting intentionally.

2. AI, Agency, and Consciousness

If we look at what is in common between the case of Eddie killing George and Parks killing his mother-in-law and attempting to murder his father-in-law is the fact that neither seemed to be aware of what it was they were doing while they were doing it. Both cases present challenges for the design of artificial agents, since it is reasonable to assume that at that time, none of what we call intelligent agents in AI are internally structured in ways that distinguish them from sleepwalkers, or that might allow them to overcome distraction during a critical moment [Bello and Bridewell, 2017]. In short, they fail to meet the criteria for being responsible for what they do because there is no meaningful distinction in such systems between awareness and unawareness of any particular piece of information at a given time. More importantly, the counterfactual statement: "I would have been aware of x at time t if only

I had done action *a* rather than having done *b*" has no real meaning unless certain (often unmet) criteria are met for a candidate AI agent, such as having an episodic memory, a mechanism for selective attention, a means to voluntarily exert control over this mechanism, the capacity to represent oneself as having causal powers, and finally the ability to draw all of these together to facilitate explanation.

The absence of these capacities in AI systems might not have been a problem in the past, when AI systems were by-and-large decision-support for human decision-makers who (possibly with help) might have done the job without AI-powered assistance. But that was then, and this is now. Present and future applications of AI are trending toward more "autonomy" for machines, with the very real possibility of a machine doing something that causes harm on the basis of information learned in an unsupervised manner, all disconnected from human minders. Who or what is the target of responsibility in these cases, or in complex cases where human−machine dyads cause harm and there is a question about how to allocate responsibility? This short paper is not the place to address such weighty matters, but the questions give us *prima facie* motivation to support building autonomous systems whose behavior is explicable in terms of mental states including consciousness/awareness, causes, exercises of agency, and so on since answering questions posed about responsibility in human−human decision-making contexts turn on deploying these concepts in judgment. A full working system is more than just a series of symbolic stand-ins for mental states: there are decisions to be made about cognitive architecture, and further decisions to be made about the mechanisms that integrate its parts in service of flexible, intelligent behavior.

3. Some Takeaways

What we have seen from both the Parks case, and the case of Eddie and George is that complex behavior is possible even in the absence or near-absence of conscious awareness. Eddie and Parks were both capable of driving vehicles in some limited way, and Parks navigated to his in-laws' home in the process of murdering his mother-in-law. We might ask what might have happened had either been aware of what they were doing. In the case of Eddie, he might have noticed the pedestrian and further noticed that the pedestrian was George. This might have altered the contents of his intention to shoot George, replacing the mode of homicide to vehicular homicide.[a] In the case of Parks, being aware of what he was doing might have given him access to background knowledge, norms, and autobiographical episodic memories of his in-laws, which by his own *post hoc* accounting were positive. This information, along with the capacity for conscious control of action, may have been enough to have saved their lives.

Taken together, the disconnection between consciously-mediated behavior and script-like automatisms, the notion that specifically conscious information widely

[a] A computational example of this scenario is forthcoming in Sec. 5.

recruits background knowledge and memories in service of deliberation, and that attention-enabling exercises of control are a good fit with a popular theory of consciousness: the Global Workspace Theory (GWT) [Baars, 1997; Dehaene *et al.*, 2006]. GWT sees cognitive architecture consisting of many distributed parts of the mind that actively shape the contents of a global workspace: a limited amount of information kept in short-term stores to be used for reasoning, decision-making, and other high-level cognitive activities. The contents of the workspace are cyclically "broadcast" throughout the mind, and coalitions among distributed processes are formed to promote their contents into the workspace in the immediate future. Parks' case was explored and explained in some detail from the perspective of GWT by Neil Levy in his 2014 book on consciousness and moral responsibility [Levy, 2014]. Levy's analysis tracks the counterfactual given above that had Parks been consciously aware, his conscious contents would have been integrated via broadcast with his dispositions and other memories. Since Levy's book is about responsibility, the conclusion he draws is that Parks isn't responsible because he was unable to recruit the background knowledge about the moral status of his actions, leaving him without a way to grasp their consequences and thus, to be responsive enough to reasons such that he might have done otherwise.

The case of Eddie killing his uncle George provides us a window onto a different aspect of the consciousness-agency nexus. The standard way of thinking about agency in philosophy is in terms of the causal theory of action. The causal theory can be understood as claiming that actions are events caused by an agent's mental states, for example, beliefs, desires and intentions. This view is the default view among philosophers, but it is also foundational in AI, yet the causal theory of action has a glaring problem: agential control. It is assumed that the causal relationship between intentions and actions are exercises of agential control, yet causal deviance examples such as the Eddie/George example vitiate the simplistic view that intentions directly cause action. Even if we were to somehow become confident in the fact that mental states, as the causal theory conceives of them, are entirely possible to implement in AI systems without the loss of any function, we would still be challenged by deviance cases, which seem to suggest that intentional action involves more than just an intention causing an action. So what is the diagnosis? Wayne Wu [2015] suggested that in not accounting for the role of attention in intentional action, the causal theory is unable to deliver an adequately explanatory concept of agential control. Wu and others defend a simple theory of attention called *selection for action*, which broadly states that if a subject **S** selects **X** to inform the performance of task **T**, then **S** attends to **X**. On this view, attention mediates the connection between having a task or intention in mind, and generating intention-consistent outcomes by keeping the intention in mind such that it becomes proximal guidance for attention. Presumably, if Eddie had been firmly focused on driving, he would have noticed the cars, street signs, and pedestrians around him. So, it would seem that Eddie's inability to exert control in order to keep

his intention to drive to George's house firmly in the front of his mind led to his unawareness of the pedestrian and eventually to the unintentional killing.

4. The ARCADIA Cognitive System

Section 3 hinted at a cognitive architecture consisting of independently operating modules that can be synchronized by the broadcasting of conscious contents and a mechanism for selective attention that interacts with intentions to produce agent-guided intentional action. The ARCADIA cognitive system was designed with consideration given to the primacy of attention in coordinating perception, cognition, and action in a way that supports the development of artificial agents in the rich sense discussed earlier in the paper [Bello and Bridewell, 2017]. Given its start as a model of attention [Bridewell and Bello, 2016], ARCADIA has been used to investigate and capture human behavior on a number of tasks that have been used to constrain theories of attention, displaying analogues to both inattentional blindness and change blindness using artificial stimuli [Bridewell and Bello, 2015]. ARCADIA has also been used to capture human performance data on various types of multiple object tracking tasks [Lovett *et al.*, 2019], as well as on memory tasks [O'Neill *et al.*, 2018], and for more complex phenomena such as combining low-level perception of numerosity with higher-level counting skills [Briggs *et al.*, 2017], as well as causal perception [Bello *et al.*, 2018].

4.1. *Representation: Components and interlingua*

ARCADIA is technically an architecture schema: in other words, it is a framework within which intelligent systems can be developed. Very simply, ARCADIA consists in a non-empty set of *components*, a *focus of attention*, and a routine for performing *focus selection*. Every component must be able to read from and write to a common data format called *interlingua*, and example of which looks like the following:

{:id 12345
:name "object1"
:arguments {:color "Red" :shape "Rectangle" :img Mat-img@34x56}
:world "working-memory"
:source working-memory
:type "instance"}.

ARCADIA is a representationally heterogeneous system: the internal processing of components can be implemented by neural networks, theorem provers, Bayesian methods, or whoever else does the job under whatever modeling constraints are in place. For cognitive modelers, these constraints may be much tighter, with the number of components and their internal processes informed by the literature in neuroscience and psychology, as contrasted with AI approaches who may be under no

other constraints but to build the most efficient algorithm possible for the task by using few components and cognitively implausible internal processes.

4.2. *The cycle*

During each ARCADIA cycle, the components will perform internal processing, framing the results as interlingua elements of the type shown above. These results are made available on the subsequent cycle to other components that might potentially use them. Because components frame results in interlingua, interoperability is ensured. And because interlingua itself is heterogeneous, and the contents of the "arguments" field is a data structure that is not necessarily representationally uniform, multiple components receiving the same bit of interlingua can work on different parts of the same structure. A component implementing a neural network can work on image data, while a reasoning component can work on symbolic elements of the same argument list.

On each cycle, a new focus of attention is selected from the set of available component outputs, and broadcasted to all components upon selection. Components are designed to be either focus-responsive or focus-unresponsive. An example of the latter might be a component meant to capture a process in early vision which we know to be unaffected by attention. Focus-responsive components will receive the broadcast, and will preferentially process the focus of attention. Fundamentally, if components are seen as simple functions that can take previous outputs of other components and process them, then ARCADIA can be seen as a simple model of computation that uses focus-selection routines to compose functions over cycle time.

4.3. *Focus selection*

How is focus selection implemented in ARCADIA? This too is up to the modeler, but the general scheme is to filter through the available content produced by components according to some scheme. In ARCADIA models, up through very recently, simple priority lists were used. An example might be a model of change detection that prioritizes the selection of any change events produced by a component to request to encode objects into working memory to focus on new objects, and so on. As the tasks ARCADIA has addressed grow more complicated, the priority list structure is being replaced by dynamically weighted lists of features to afford more flexibility, although this is outside the scope of our discussion.

4.4. *Intentions, tasks, and task-switching*

Every ARCADIA model must have a focus selection routine called the default attentional strategy, however, most ARCADIA models are models of tasks, and thus have internal representations of task structure. Task representations in ARCADIA are distributed across the system and consist of semantic information about the task,

such as its name, along with procedural information in the form of so-called stimulus-response (SR) links that match against the collection of contents produced on each cycle by the model's components, and generate action requests [Bridewell *et al.*, 2018]. Along with this standard set of task-related information, each task representation comes along with its own set of attentional priorities that attunes the model to task relevant features of the environment. When the model adopts an intention to X, the task representation for X is loaded into working memory, and the attentional priorities that are part of the task representation for X are thereby used for focus selection until the intention is dropped and a newly adopted intention leads to different task representations being recruited and their corresponding attentional strategies being used to guide attentional focus.

4.5. *Relevance to attention-guided intentional action*

Recall that the causal deviance case involved acting while distracted. Eddie was still driving his vehicle while ruminating on killing George, keeping his focus off of noticing relevant aspects of the environment, including the presence of pedestrians. First, one should note Eddie's ability to drive, however poorly, while his attention is fully consumed by rumination, suggests that driving is at least partially served by the sort of automatic behavior that we saw in the case of Kenneth Parks. Second, if we construe rumination as a task of its own, it seems as if its attentional priorities compete with those required to drive safely, and that they cannot be simultaneously driving attentional focus. Periodically switching back and forth between ruminating and safe driving might have been sufficient for avoiding the collision with the pedestrian in Eddie's case, but the vignette assumes Eddie's attention is completely preoccupied with his thoughts in a way that prevents multitasking of this sort. Since some ARCADIA components are not necessarily responsive to focus broadcast, they can be used to model automatic behavior. And since ARCADIA tasks come along with their own attentional priorities that dominate attentional selection if the task is selected as the system's active intention, we can capture differences in how attention is allocated over time for both rumination, which prioritizes attention to internally-generated speech and imagery, and driving, which might prioritize attention to the presence of pedestrians, signs, cars, and road-lines. Because of the aforementioned competition between priorities, when the system's active intention is to ruminate, its attentional focus will be on inner speech or imagery if any is being produced by other components, rather than being on any driving-related items.

5. An Example of Attention-Mediated Intentional Action in ARCADIA

To illustrate the difference between intentional and unintentional action, we implemented a version of the Eddie/George case in ARCADIA and modified the scenario further to illustrate intentional action. These models comprise close to 40

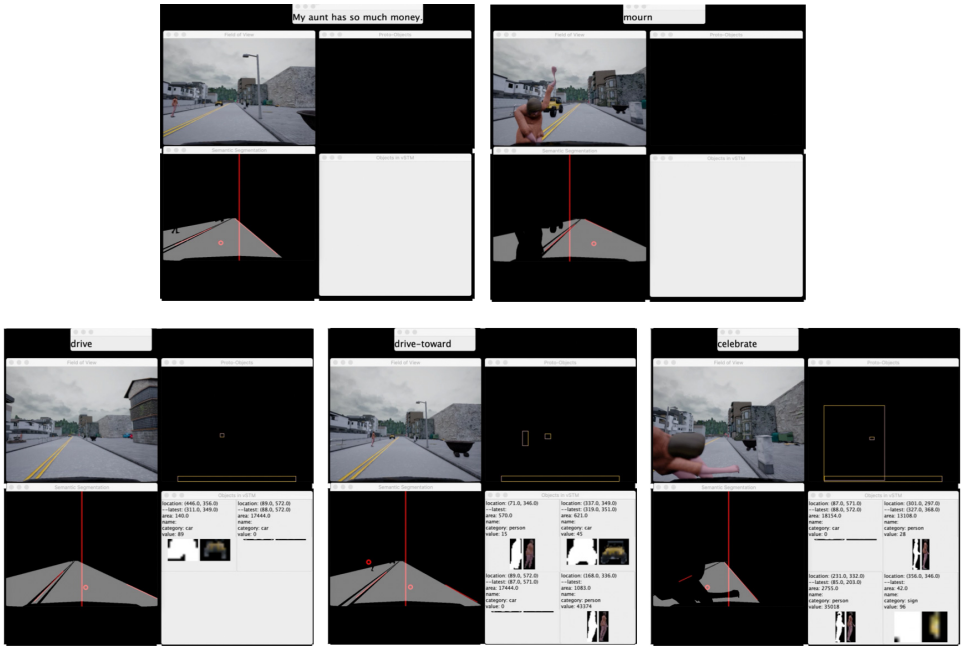

Fig. 1. Top: Unintentional killing. The model relies on a single point in the periphery to control driving. Attention is focused on the contents of inner speech, and so visual memory (lower right corner of each video) is empty. Bottom: Intentional killing. The model is focused on driving and atttentional priorities include cars, signs, and pedestrians as evidenced by the contents of visual memory. Target is identified and marked with a second control point for steering, and a trajectory is planned to run the target down.

interacting components, including functionality for visual processing, short, and mid-term memory stores.

As we mentioned earlier, some of ARCADIA's components are not responsive to the focus of attention. In the models shown in Fig. 1, two simple PID controllers from the throttle and steering wheel produce outputs on every cycle subject to the position of one or more control points (the red dots shown in the lower-left windows). The bottom-most control point is automatically computed and is available to the system without having to attend, however steering based only on this one control point results in erratic driving that fails to take upcoming information into account in the control computations. A second control point can be placed on the road ahead or on any other person or object in the visual field to further inform steering or in the case of the intentional killing sequence in the bottom row, to mark a target to aim the vehicle towards.

In the top row of the figure, we see ARCADIA focused solely on the contents of inner speech, and relying solely on attention-free one-control-point driving behavior. The driving is erratic, and the car runs down the pedestrian by accident, followed by mourning. Because the attentional strategy associated with rumination prioritizes focus on interlingua elements containing inner speech at the expense of everything

else, no information from the outside world ends up in visual memory (the lower-right corners of each window), meaning that ARCADIA never "saw" the pedestrian at all. The bottom row is a contrast, where ARCADIA is focused on driving (see the small "active intention" indicator at the top of each window), and driving-related attentional priorities drive focus toward cars, pedestrians, and signs. Once the pedestrian is identified as the target, a second control point is placed to inform steering and the car is driven into the target, followed by a celebration of success.

6. Discussion

We have argued throughout that a model of cognitive architecture grounded in what we know about human cognitive architecture and especially about attention is a productive research strategy given likely future demand for AI capable of rich human−machine interaction. It seems reasonable enough to assume that this kind of interaction will depend on machines that not only have the capacity to explain themselves to humans, but also to be bonafide agents in the first place, capable of acting intentionally, and exerting (self-) control when required. We have seen from both case studies we examined, that human agency exists within an overall cognitive architecture that is distributed, locally inflexible at the level of modules, and is reliant on mechanisms like attention that prioritize one item at the expense of potentially many others.

On the face of it, this is antithetical to the program of optimality that is typical of AI agent building, whether by way of reinforcement learning, or symbolic planning. The same can be said for most contemporary approaches to perception and action in artificial agents and robots. While there is evidence that things are changing slowly, with mechanisms of "attention" being added to deep neural architectures and yielding great success, historically, whole-scene and whole-plan space processing was the norm. It may seem now that advances in both the speed and amount of computation able to be performed by a system might obviate mechanisms like attention, but caution ought to be taken here. Sebastian Watzl draws an excellent analogy to our human lives before and after the arrival of Google search [Watzl, 2017]. Nothing has changed about our need to impose priorities on tour information gathering activities. The amount of information we have access to, and the speed at which it can be accessed doesn't change the fact of the matter. This is, in very broad strokes, one of attention's primary functions according to Watzl, Wu, and others. It is in concordance with how attention operates in ARCADIA. The study of machine consciousness and conscious agency will benefit greatly by systems-level accounts of the relationship between attention, intention, and action.

Acknowledgment

This work was kindly supported by Grant Nos. N0001419WX00023 and N0001419WX00020 from the Office of Naval Research. The opinions expressed

herein are solely those of the authors and should not be taken to express any policy or position of the Department of Defense of the United States government.

References

Baars, B. J. [1997] In the theatre of consciousness. Global Workspace Theory, a rigorous scientific theory of consciousness, *J. Conscious. Stud.* **4**, 292–309, http://www.ingenta-connect.com/content/imp/jcs/1997/00000004/00000004/776.

Bello, P. and Bridewell, W. [2017] There is no agency without attention, *AI Mag.* **38**(4), 27–34.

Bello, P., Lovett, A., Briggs, G. and Neill, K. O. [2018] An attention-driven computational model of human causal reasoning, in *Proc. 40th Annual Meeting of the Cognitive Science Society*, (Madison, WI), pp. 1353–1358.

Bridewell, W. and Bello, P. F. [2015] Incremental object perception in an attention-driven cognitive architecture, in *Proc. 37th Annual Meeting of the Cognitive Science Society*, 279–284.

Bridewell, W. and Bello, P. [2016] A theory of attention for cognitive systems, in *Fourth Annual Conf. Advances in Cognitive Systems* (Evanston, IL), pp. 1–16.

Bridewell, W., Wasylyshyn, C. and Bello, P. [2018] Towards an attention-driven model of task switching, *Adv. Cognitive Syst.* **6**, 1–6.

Briggs, G., Bridewell, W. and Bello, P. F. [2017] A computational model of the role of attention in subitizing and enumeration, in *Proc. 39th Annual Meeting of the Cognitive Science Society*, (London, UK), pp. 1672–1677.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J. and Sergent, C. [2006] Conscious, preconscious, and subliminal processing: A testable taxonomy, *Trends Cognitive Sci.* **10**(5), 204–211, doi: 10.1016/j.tics.2006.03.007, http://www.sciencedirect.com/science/article/pii/S1364661306000799.

Levy, N. [2014] *Consciousness and Moral Responsibility* (Oxford University Press).

Lovett, A., Bridewell, W. and Bello, P. [2019] Selection enables enhancement: An integrated model of object tracking, *J. Vis.* **19**(14), 23, doi: 10.1167/19.14.23.

O'Neill, K., Bridewell, W. and Bello, P. [2018] Time-based resource sharing in ARCADIA, in *Proc. 40th Annual Meeting of the Cognitive Science Society*, (Madison, WI), pp. 828–833.

Watzl, S. [2017] *Structuring Mind. The Nature of Attention and How it Shapes Consciousness* (Oxford University Press).

Wu, W. [2015] Experts and Deviants: The story of agentive control, doi:10.1111/phpr.12170.