Artificial Agency Requires Attention: The Case of Intentional Action [⋆]

Paul Bello, Kevin O'Neill & Will Bridewell

Naval Research Laboratory, Washington DC 20375, USA
{paul.bello, kevin.oneill, will.bridewell}@nrl.navy.mil

Abstract. What does it mean to build an *artificial agent*? In prior work, we have argued at length that the promise of AI introduced into our social milieu calls for a rethinking of the question [1]. We have claimed that agency of the form that all of us naturally recognize requires consciousness to support reasons-responsive choice (ibid). In this short paper, we further claim that consciousness, or something close enough, is often necessary for intentional action as well, and explore the connection between them through thought experiments and computational modeling.

Keywords: Attention · Intentional Action · Artificial Agency.

1 Motivation

Imagine an entitled nephew whom after many years of expecting to inherit his wealthy uncle's fortune comes to learn that he has been cut out of his uncle's will. Infuriated, he puts his rifle in his car, hops in, and races toward his uncle's house. On the way he is consumed with thoughts of shooting his uncle and the resultant bloody carnage. With a *thud*, he is broken out of his fantasy as his car careens into an unnoticed pedestrian. Upon getting out of the car he realizes he has run down and killed none other than his uncle! The question then is whether the killing was intentional[1]. Let us re-run the thought experiment, but allow for the fact that something salient in the environment re-centers the nephew on driving such that the pedestrian is identified prior to impact and specifically identified as the nephew's uncle. At this point, the nephew decides to modify his intention to shoot his uncle by merely running him over instead. Say he does so. Did the nephew kill his uncle intentionally in this version of the scenario?

In the case of the angry nephew, intuition suggests that the first variant is unintentional killing while the second variant is textbook intentional killing. The inference to intentionality hinges on the reader's knowledge that the nephew was distracted in the first variant and thus unaware of what he was doing while it was

[⋆] The authors would like to acknowledge generous support from the Office of Naval Research under Award N0001418WX00166. The views expressed in this paper are solely the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

[1] This example is a variant on the original thought experiment presented in [4]

happening. Lack of awareness is one critical factor that can undermine *agentive control*, or the ability of the agent to appropriately guide action.

2 Attention and Intentional Action

Rather than wade directly into the murky waters of consciousness, we have chosen to start with *attention*, which has a close relationship to conscious awareness [5]. To begin giving a computational account of intentional action we draw upon work by the philosopher Wayne Wu, who situates attention as mediating the relationship between having an intention and successful intentional action [7]. Doing so makes room for *agentive control*, such that an agent merely having the appropriate beliefs, desires, and intentions isn't sufficient in all cases for successful intentional action. Rather, attention must be paid in appropriate measure to the task at hand to intentionally succeed at achieving one's goals, as was demonstrated in the case of the angry nephew.

Elaborating on work by Wu, we have developed a novel computational treatment of intentions that treat them as being partially constituted by attentional priorities in the ARCADIA computational cognitive system [3]. Each intention consist of high-level semantic information describing its conceptual content (i.e., a linguistic description of the intended action), the set of attentional priorities mentioned above, and a set of stimulus-response (S-R) links which are the procedural knowledge used to execute the intended task. When an intention is adopted it is loaded into task memory. While task memory can accommodate more than one intention to facilitate multitasking and even modest concurrency in execution, only one intention at a time is dominant. An intention is dominant in task memory just in case its attentional priorities are currently guiding the system's focus of attention. Keeping an intention dominant is a matter of periodically focusing on it. We have implemented a facility for inner speech in the system, such that when the name of an intention, or semantically-related content is subvocalized, the intention remains active. Switching intentions is a matter of a non-dominant intention receiving focus, and having its attentional priorities guide system behavior.

Insofar as attentional priorities guide the focus to external features, regions, objects, and events, they largely determine the content of ARCADIA's short-term memory stores. Items in short-term memories roughly correspond to what Ned Block famously labeled *access conscious* contents, which he characterized as being poised to be used in the rational control of thought and action [2]. Taken together, dominant intentions in ARCADIA partially structure (access) consciousness in virtue of their connection to attention. We have put all of this functionality to work in developing a simple agent corresponding to the angry nephew in the scenarios detailed earlier. Initially, the agent has three intentions: to plan out the specifics of the murder, to drive to his uncles home, and finally to kill his uncle. Because ARCADIA supports multiple active intentions, the agent is able to drive (albeit poorly), and plan at the same time. The S-R links associated with driving respond directly to aspects of the environment that are processed outside of the focus of attention up to the level of coarse-grained semantics. In the first variant, planning dominates and the agent neglects to periodically switch back to focusing on driving, causing the agent to unintentionally run over what turns out to be its uncle. In the second variant, the agent is similarly focused on planning, but a shiny red stop-sign grabs the agent's attention, facilitating a switch back to driving. Once the agent is focused on driving, the uncle is spotted. Recognizing the uncle affords the agent the opportunity to fill out the plan by using the car to run the uncle over rather than either shooting or stabbing him, which were the two options under prior consideration. Once planned, the agent focuses once more on driving and runs the uncle over straightaway, killing him intentionally.

3 Final Thoughts

Along with others [6, 5], we recognize that any attempts to computationally model what we ordinarily think of as consciousness will need to grapple with perception and attention. Of course, one might imagine building a purely Cartesian AI, but such a creature would have no appreciable connection to the external world. Ultimately, the type of agent we have in mind exists in the world, perceiving continuously and acting often. Selective attention structures both access consciousness and phenomenal consciousness: when we move our eyes, peripheral objects come into detailed focus, and the details of objects previously in focus fade to some degree, modulo whatever detail may have been encoded in short-term memory. Understanding and modeling conscious cognition is ultimately a systems-level enterprise, involving sensation, perception, attention, memory, inference, and action. While an enormous amount of empirical work remains in clarifying the relationships between these various capacities with respect to consciousness, it seems undeniably true that a rich model of attention is at their heart, and required for the kind of conscious agency that is of interest to all of us who look to a future where machines and man are bound by the same social and cultural mores [1].

References

1. Bello, P., Bridewell, W.: There is no agency without attention. AI Magazine **38**(4) (2017)
2. Block, N.: On a Confusion About a Function of Consciousness. Brain and Behavioral Sciences **18**(2), 227–247 (1995)
3. Bridewell, W., Bello, P.: A theory of attention for cognitive systems. In: Fourth Annual Conference on Advances in Cognitive Systems. vol. 4, pp. 1–16. Evanston, IL (2016)
4. Chisholm, R.: Freedom and Action. In: Lehrer, K. (ed.) Freedom and Determinism. Random House (1966)
5. Prinz, J.: The Conscious Brain. Oup Usa (2012)
6. Watzl, S.: Structuring Mind: The Nature of Attention and How It Shapes Consciousness. Oxford University Press (2017)

7. Wu, W.: Experts and Deviants: The Story of Agentive Control (2015). https://doi.org/10.1111/phpr.12170