Extended Control Systems: A Theory and its Implications

Hunter R. Gentry
hgentry@wisc.edu
University of Wisconsin-Madison

[PREPRINT] Forthcoming in *Philosophical Psychology;* accepted 06/21/2020

Abstract: Philosophers and cognitive scientists alike have recently been interested in whether cognition extends beyond the boundaries of skin and skull and into the environment. However, the extended cognition hypothesis has suffered many objections over the past few decades. In this paper, I explore the option of control extending beyond the human boundary. My aim is to convince the reader of three things: (i) that control can be implemented in artifacts, (ii) that humans and artifacts can form extended control systems, and (iii) that perhaps extended control ought to be preferred over extended cognition. Using the objections to extended cognition as constraints on my own extended theorizing and the example of autofocus systems in cameras, I decompose and localize the components of an autofocus system that realize the central properties of control from a plausible theory of control in the literature. I then provide criteria according to which control can be extended in a system. Finally, I consider how this theory of extended control ought to be preferred to theories of extended cognition.

Introduction

Recently, philosophers of mind and psychology have highlighted the concept of control as a neglected, but central component of agentive action (Bermúdez, 2017; Buskell, 2015; Fridland, 2014, 2017, 2019; Shepherd, 2015a; Wu, 2016). Control has been thought important to a wide variety of other major concepts such as the distinction between knowing-how vs. knowing-that (Fridland, 2015; Pavese, 2017), skill and expertise (Bermúdez, 2017; Christensen, Sutton, and McIlwain, 2016; Fridland, 2014; Wu, 2016), and even normatively-thick concepts like justification (Pavese, 2016) and moral responsibility (Fischer and Ravizza, 2000; Shepherd, 2014, 2015b). In the field of control systems engineering, researchers are concerned to model machine functions engaged in controlled processing. Here, a foundational assumption is that artifacts engage in controlled processing over the environment to bring about certain states of affairs. For example, a thermostat controls for the temperature in the room by monitoring the

temperature continuously and causing the air conditioner to turn on when the temperature moves out of the desired range. Or in the field of human-robot interaction (HRI) (for review see Bauer et al., 2008) control is central to understanding complex interactions between humans and robots in goal-directed behavior. Researchers have offered various models corresponding to various levels of automation (see Sheridan, 2011). For example, supervisory control (Sheridan,1992; 2011) is, "the idea is that a human supervisor instructs and gets feedback through an intermediary computer which *itself* closes a direct control loop through an artificial measurement means and a feedback-controlled process" (pg. 663, emphasis added).

Advancements in technology have relieved humans of all kinds of tasks-- calculators calculate mathematical functions so that we do not have to use pen and paper or mental math, cruise control controls the speed of the car so drivers do not have to do it themselves. These instances of relief, or "cognitive offloading", have led some philosophers and cognitive scientists to wonder if cognition extends out into the world (Clark & Chalmers, 1998; Haugeland, 1998; Hutchins, 1995; Rowlands, 2010). In general, there are four ways to talk about extracranial cognition: embodied, embedded, extended, and enactive.

> **Embedded Cognition**: Cognition causally depends upon bodily and/or environmental processes.
> **Embodied Cognition**: Cognition is partially constituted by processes in the body (that are not in the brain).
> **Extended Cognition**: Cognition is partially constituted by processes in the body (that are not in the brain) and the environment.
> **Enactive Cognition**: Cognition is partially constituted by the ability or disposition to act (Newen, de Bruin, and Gallagher, 2018).

Debates between these camps concern where cognition is located. Embedded theorists (that are also not embodied, extended, or enactive theorists) (Adams & Aizawa, 2010; Rupert, 2009) tend

to be more traditionalist in that they think cognitive processing is only in the head. Embedded theorists give various reasons for thinking this, but they all provide some necessary condition on a process' being cognitive that rules out its extension. For example, Adams & Aizawa (2010) claim that cognition necessarily involves original, or non-derived, content. Because artifacts and the environment do not carry original content, cognitive processing only occurs in the head. In general, embedded theorists provide two objections to extended cognition: (1) the mark of the cognitive and (2) the coupling-constitution fallacy. The mark of the cognitive objection is the idea that extended theorists must provide a theory of cognition that clearly marks its bounds *prior* to claims of extension. In other words, extended theorists must answer the "what?" question about cognition before answering the "where?" question. Relatedly, the coupling-constitution fallacy states that extended theorists cannot infer a constitutional claim from a causal claim. For this inference to be valid, extended theorists need another premise. As the reader will see later, that extra premise comes out of a theory of cognition. Of course, these critiques are aimed at extended cognition theorists, but they apply to any theory that attempts to show that some mental property extends.

Josh Shepherd (2014), a philosopher working on control, has granted that *control* might extend beyond the boundary of humans, but little attention has been paid to how exactly this might happen. In this paper, I argue for a theory of extended *control.* The objections to the extended cognition hypothesis listed above, I take it, can be generalized to apply to any theory that claims some mental property extends. So, I use these objections as constraints on my theorizing about control. I will be arguing for the claim that control is partially constituted by processes in the body (that are not in the brain) and the environment. That is to say that control

spans the brain, body, and environment. To overcome the mark of *control* objection, I offer a plausible theory of control (Shepherd, 2014). I then use that theory to guide mechanistic analysis of an artifact to show that the artifact implements controlled processes.

Upon establishing that artifacts can be controllers, I turn to a discussion on the formation of extended control systems between agent and artifact. I provide criteria under which control extends in a system. These criteria include, information processing, expectation, and delegation. For control to be extended, the relevant artifact must process information about the controllee to maintain the control autonomously. Expectation involves the agent expecting the artifact to deliver certain outputs-- those appropriate to the completion of the task. Finally, the agent delegates the subtask to the artifact, where delegation denotes the introduction of the artifactual contribution by the agent. The upshot of this account is two-fold. First, extended control potentially provides a framework for understanding responsibility in the moral and/or epistemic domains.[1] The other is that we might have reason to prefer extended control over extended cognition if extended control can accomplish the explanatory goals of extended cognition while avoiding the traditional objections.

The strategy for the paper is as follows: in section 1, I develop constraints on extended theorizing that come out of objections to the hypothesis of extended cognition. Hence, in section 2, I will present a plausible theory of control from the work of Josh Shepherd (2014). In section 3, I will present my argument by mechanisms that will serve to establish that artifacts can implement control. Section 4 will serve to flesh out the theory by describing some criteria that I

[1] This ultimately depends on whether you think control is a thick or thin notion. If thick, then control is morally, legally, and/or epistemically loaded-- that is, the notion bears on these debates. If thin, then control does not bear on these debates at all. I am remaining agnostic on this point.

take to be strengthening the lower bounds of extended control.[2] This section should block the criticism that extended control is too permissive, while also clarifying key concepts in how control can be extended. Finally, in section 5 I consider whether extended control creates a consequence for extended cognition. I close by considering avenues for future research.

1.0 Constraints on Extended Theorizing

To begin, I want to consider two cases that I take are paradigmatic of extended control. Then, I want to consider some objections to extended cognition theorizing that I will develop into constraints. These constraints serve as a guide to extended control theorizing.

1.1 Two Cases of Extended Control

Having an ecumenical basis of control will be useful at this point. Consider Dennett's (2015) explication of the term:

> A controls B if and only if the relation between A and B is such that A can drive B into whichever of B's normal range of states A wants B to be in. (If B is capable of being in some state s and A wants B to be in s, but has no way of putting B in s, or making B go into s, then A's desire is frustrated, and to that extent A does not control B.) (Dennett, 2015, pg.57).[3]

Dennett makes clear that "wants" and "desires" need not be interpreted as human states. Perhaps a better way to put it is goal-states-- "A has the goal that B move into state s". To be a controller is, in part, to have these goals about some other thing. For something to be a controllee is for it to have a range of states that it can move into.

[2] I don't take myself to be offering necessary and sufficient conditions, but rather describing the dimensions of extended control systems, particularly the minimal benchmarks of such a system to form.

[3] I am not suggesting an intentional stance view towards systems by using Dennett here. Moreover, I should note the distinction between Cummins and Millikan functions. I am committed to teleological notion of function, not mere efficient causes.

The following cases are intended to clear intuitions about where control can be located. I aim to open the possibility that control can be shared between an agent and artifact. Consider first a case of cruise control in a car.

> *Cruise Control:* Caroline is driving to her parents house on a long, empty stretch of highway. The speed limit is 65 miles per hour on this road and she knows that the cops patrol it frequently. She'll be on this highway for at least an hour and does not want to have to constantly monitor her speed. She remembers that her car is equipped with cruise control and thus engages it.

Now consider a photographer using a camera with autofocus:

> *Autofocus:* A sports photographer prepares his camera for the 100m dash while standing on the sideline. As the runners take their mark, the photographer takes aim. The gun cracks and the runners are off. It's over in less than 15 seconds, but the camera focuses the lens automatically-- he just points and shoots. The photographer gets many clear photos of the runners.

To put these two cases into a context, consider the alternative (manual) modes of accomplishing the relevant tasks. Caroline would manually control the speed using cognition and motor control and the photographer would have to manually focus the lens of his camera. These artifacts assist in goal-directed behaviors by relieving some of the work that the agent would have done. Seemingly, cruise control systems and autofocus systems relieve the agent by controlling for features of the relevant task. Thus, functionally, these systems play the same role that, say, motor control plays in carrying out a task. Motor control controls for features of the task, namely, moving the limbs thus and so. Just the same, these artifacts control for features of the task.

Using Dennett's terms, we can describe *Autofocus* as follows: let A be the autofocus subsystem of the camera and let B be the lenses. A has the goal that B be in focus relative to the subject. Autofocus subsystems exert causal influence on lenses to drive them into particular

states that satisfy the goal of bringing the subject into focus. So, the autofocus subsystem controls for focusing the image. But we can further describe the case. The autofocus subsystem is dependent upon the photographer for its exerting causal influence on the lenses. Let A* be the photographer and B* be the autofocus subsystem. Thus, we can say that A* has the desire (or intention, or goal) of B*'s being in the state of "on" such that it can exert causal influence on the lenses. The photographer, using his psychological control mechanisms, drives the autofocus subsystem into a state that enables it to drive the lenses into a state that focuses the image. So, the photographer controls the autofocus subsystem which, in turn, controls the lenses.

1.2 Criticisms of Extended Mentality as Constraints

Adams and Aizawa (2010) have argued against the hypothesis of extended cognition by claiming that extended theorists do not have a theory of cognition that captures its bounds, i.e., makes a distinction between what is cognitive and what is not. For them, a cognitive process is, at least in part, characterized by how the process works and/or what mechanism(s) realize it and involves non-derived representations.[4] In contrast, Haugeland's theory of systems entails that systems are defined by what interfaces with what. In other words, we can make a system intelligible by analyzing merely the connections between the component parts. Adams and Aizawa claim that this definition is insufficient. That is, systems generally, and biological systems in particular, are defined by their *structure* and function. Thus, a constraint upon extended therozing is that one's theory, of whatever property is being extended, must make

[4] I should emphasize here that the derived/non-derived representation distinction is fairly loaded from the mental representation literature and that this distinction does a lot of work for Adams and Aizawa. See Huebner (2014, pp. 169-82) for an argument against the claim that mental representations in cogniziers are necessarily or essentially non-derived.

reference to the nature of the underlying processes realizing said property, i.e., taxonomy-by-mechanism.

Rupert (2009) argues along similar lines that the extended theorist must have a principle of demarcation--demarcating the bounds of whatever property is claimed to extend. In the case of control, Rupert's adapted proposal is the following:

> A state instantiates [control] iff it consists in, or is realized by, the activation of one or more mechanisms that are elements of the integrated set members of which contribute causally and distinctively to the production of [controlled behaviors] (pg.42, adapted).

Adams & Aizawa and Rupert demand this kind of constraint mainly for explanatory purposes in cognitive science. Consider the following example from Adams and Aizawa:

> The set of muscles in the human body constitutes the muscular system. In general, the muscles in the muscular system do not interface with each other; hence they do not constitute a system in Haugeland's sense. Even antagonistic muscles, such as the lateral and medial rectus muscles of the eye, which move the eye left and right, do not connect to each other. But, even if one were to say that antagonistic muscles have a kind of indirect interface, there are other combinations of muscles that do not stand in such antagonistic relations. The lateral and medial rectus muscles of the eye do not, for example, interface in any natural way with the gastrocnemius muscles of the calves. What appears to unify the muscles of the body into a system appears to be a commonality of function and a commonality of underlying mechanism (2010, pg.115).

I take this example to be suggesting that a purely functional analysis of the system under consideration is not sufficient to establish that cognition (or whatever the relevant property is) extends. That is, there must be at least some attention paid to the underlying mechanisms that realize the relevant property, especially to explain the success of cognitive science which depends upon persisting systems. The persisting system, for Adams & Aizawa and Rupert, is the intracranial mechanisms that realize cognitive processes. To be clear though, these philosophers are not suggesting that cognition must contain exactly the mechanisms found in humans. They leave it open that cognition might be realized by mechanisms other than humans, but as a matter

of empirical fact, it is not. Thus, for my purposes, it is also open that control can be realized by other mechanisms besides those found in humans. I contend that there are such cases.

Proponents of extended cognition claim that cognition extends into the surrounding environment of the agent through being coupled to the agent. The coupling is what makes the relevant part of the environment part of the cognitive system for the time being. Adams & Aizawa (2010) argue that extended theorists have committed a fallacy in moving from a causal relation to a constitution relation-- the coupling-constitution fallacy (CC fallacy). Consider Otto and his notebook (Clark and Chalmers, 1998). Here, Otto is coupled to his notebook as he relies upon it to "remember" where the MOMA is. But, according to the CC fallacy, it does not follow from this causal relation that the notebook is a part of Otto's cognitive system. The fallacy then is to move from a claim about the causal coupling of some part of the environment to a cognitive agent to the conclusion that that the part of the environment is a part of the cognitive agent or cognitive processing. In other words, the fallacy is to confuse causal relations with part-whole relations-- the former being a diachronic relation whereas the latter is synchronic.

The CC fallacy, if it is a fallacy, is a rather obvious mistake and so one might wonder how it is that smart philosophers and cognitive scientists could make it. Adams & Aizawa provide an explanation:

> [I]f the fact that an object or process X is coupled to a cognitive agent does not entail that X is part of the cognitive agent's cognitive apparatus, what does? *The nature of X of course*. One needs a theory of what makes a process a cognitive process. One needs a theory of the "mark of the cognitive" It won't do simply to say that a cognitive process is one that is coupled to a cognitive agent, since this only pushes back the question. One still needs a theory of what makes something a cognitive agent (Adams and Aizawa, 2010, pg. 68; my italics).

Adams & Aizawa are offering an error theory for why one might commit the CC fallacy-- the idea here is that extended theorists do not have a theory of the cognitive. One needs a theory of the cognitive to deliver verdicts about what parts of the environment are cognitive prior to claims of extension.[5] In other words, it is illicit to move directly from a causal claim to a constitution claim because one needs another premise in the argument-- namely, a premise derived from the verdict that a theory of the cognitive delivers about the relevant part of the environment.

Interim conclusion: extended theorists must provide a theory for whatever property is claimed to extend. This theory is supposed to guide mechanistic analysis of a system that will deliver a verdict about whether the system's components implement the property in question. The theory allows for extended theorists to then validly infer from a causal claim to a constitution claim by supplying a missing premise. In the next sections, I intend to meet these constraints for the property of control.

2. A Theory of Control

One constraint on extended theorizing that comes out of the critiques of extended cognition is that fans of extended cognition need to provide a theory of cognition that clearly marks its bounds.[6] In the case of extended control then, I need to provide a theory of control that marks *its* bounds. In this section, I will be relying upon a plausible theory of control from Josh Shepherd (2014) to tease out the properties that are most central to the notion of control.[7]

[5] See Rowlands (2009) for an argument that the CC fallacy reduces to the mark of the cognitive objection.
[6] It is notable that some extended cognition theorists think that their theory goes some way towards understanding cognition, but Adams & Aizawa and Rupert criticize them for not adequately specifying the property they are (extended) theorizing about. Control has an upper hand here because, for the most part, it is well understood in areas such as cybernetics, engineering, and neuroscience (especially on motor control). Or at least, control is better understood than cognition.
[7] I realize that there are other theories of control available, but I do not have the space to consider alternatives. The fact that my view is supported by this major theory in the literature, suffices for my purposes.

2.1 Shepherd's "Contours of Control"

Shepherd (2014) provides a clear and general account of control in intentional action. While his account focuses on agents' implementation of control, it seems to me this is only because he talks in terms of intentions. As we saw on Dennett's account, control need not be pitched in those terms. Indeed, Shepherd seems to acknowledge as much in the following quote:

> In my view, it is plausible to extend an agent's physical constitution... to relevant tools in the agent's environment (Shepherd, 2014, pg.404, footnote 8).[8]

In this respect, we might slightly revise his view to account for this as follows:

> An agent J exercises control in service of [a goal (or derived intention)] I to degree D in some token circumstance T if and only if (a) J's behavior in T approximates the representational content of I to (at least) degree D, (b) J's behavior in T is within a normal range for J, where the normal range is defined by J's behavior across a sufficiently large and well-selected set of counterfactual circumstances C of which T is a member, (c) the causal pathway producing J's behavior in T is among those normally responsible for producing J's successes at reaching the level of content-approximation represented by D across C (Shepherd, 2014, pg.410).

Let us consider each condition in turn. (a) seems to be an accuracy condition on the representation of the goal to the corresponding behavior. This matters to control, says Shepherd, because a creature who has a goal state to, say, take clear photos, but the behavior implemented to achieve this goal does not achieve the end, is not in control of the behavior. The behavior ought to be appropriately guided to the desired end. Control seems to directly vary with the attainment of the goal. Shepherd invokes an approximation modifier here because there are clearly cases where the goal is not satisfied (the representation does not match the behavioral output), but the creature is still in control. This happens when, say, a professional basketball

[8] Thanks to an anonymous reviewer for pointing this out to me.

player shoots a three pointer, but just barely misses. The player is still in control despite not making the shot.

Condition (b), Shepherd tells us, is meant to capture flexible repeatability. We would like to think that a controller can repeat the relevant task in various circumstances including circumstances that involve environmental perturbations. A baseball player that can throw an 80 mph curveball in perfect conditions is impressive and might give us evidence that he has control over his behavior. But better evidence is when he can throw that 80 mph curveball in various kinds of environmental conditions, e.g., in the rain, in the snow, with bases loaded, etc. However large the set of counterfactual circumstances contributes to the degree of control one has relative to that behavior.

Finally, condition (c) is meant to exclude deviantly caused behaviors. An example of a deviantly caused behavior would be where a basketball player intends to make a shot, throws the ball, and a large gust of wind carries the ball into the basket. Were the gust of wind not present, then the ball would not have gone into the basket. Here, there is a perfect match between the agent's represented intention and the outcome, but the outcome was not due to the causal powers of the agent. This does not strike one as a case of a high degree of control. Put another way, suppose a different basketball player throws the ball and it rolls around the rim, but does not go in. There is not a perfect match between the represented intention and the outcome, but it's very close. This constitutes a high degree of control, but not as much as the other basketball player. And what makes this difference is a gust of wind-- not factors attributable to the agent. This is unacceptable, hence condition (c).

So on Shepherd's view of control, the degree of control a controller has is proportional to the degree that the representation of the goal (or derived intention) matches the outcome, how large the set of counterfactual circumstances is, and the behavior must be non-deviantly caused. In the next subsection, I will turn to extracting general properties of this theory to guide mechanistic analysis of an artifact that seems to implement control.

2.3 Properties of Control

Shepherd's account is supposed to capture how humans implement controlled sequences of actions to achieve ends. But we can extract some general properties of control that, as I will argue, non-human systems can instantiate. In engineering control theory, there is a distinction between open loop systems and closed loop systems. The basic idea is that open loop systems do not receive feedback about the state of the controllee, while closed loop systems do. For example, a standard toaster, once set to toast bread, cooks until the timer goes off. This is an open loop system. Were the toaster to receive feedback about the state of the bread while cooking, say, how dark the bread is, then it would be a closed loop system. The advantage that closed loop systems have over open loop systems is that they are more flexible and less prone to error (assuming the feedback mechanism is not malfunctioning).

Shepherd's condition (a) insists on a (approximate) correspondence between the goal state and the outcome-- a perfect match yields a high degree of control while deviance lowers the degree of control. Closed loop systems, in virtue of receiving feedback about the state of the controllee, are in a better position to get a perfect match between goal and outcome. That is, they can appropriately guide the behavior (to a greater extent than open loop systems) to the desired outcome. Furthermore, these systems are in a better position to repeatedly get a perfect match in

various counterfactual circumstances. That is, closed loop systems satisfy condition (b) for flexible repeatability as well. The imagined closed loop toaster receives constant feedback about the state of the cooking bread such that the toaster can update proximal goals to converge on the desired distal goal. The information it receives about the bread would include the relevant influence from the environment on the bread as well. In this sense, closed loop systems yield a higher degree of control than their open loop counterparts by processing relevant information about the controllee. To be sure, closed loop systems yield a higher degree of control by processing information about the controllee's current state which allows for flexible repeatability and appropriate behavior guidance.

There are three crucial properties of a controlled process: guiding action appropriately, repeatable flexibility, and information processing. Importantly, for a system to guide action appropriately, it must be flexible; and to be flexible, it must process information. In the next section, I will use these properties to give an argument from mechanisms that is intended to show that, as a matter of fact, artifacts can, and do implement controlled processes.

3.0 Mechanisms of Autofocus Subsystems and the Realization of Control

To show that control is implemented by other mechanisms besides those in humans, I need to give at least one example wherein an artifact, in virtue of the structure and function of its underlying mechanisms, implements the properties I identified from Shepherd's account of control: guiding behavior appropriately, flexible repeatability, and information processing about the controllee. To show this, I will describe the mechanisms at work in a camera with an autofocus subsystem (specifically a DSLR) and then argue that these mechanisms do, in fact, implement control by instantiating the identified properties.

In a standard DSLR camera, there are a few components necessary to take photos.[9] There are: lens(es), reflex mirror, sub-mirror, image sensor, phase detection sensor, pentaprism, and viewfinder. In general, light passes through the lens(es) and enters the camera. It then travels to the back wall of the camera where it is reflected off the reflex mirror into the pentaprism where it is then reflected twice and enters the viewfinder. Here is where the photographer can see the image captured by the light, i.e., the "through-lens" image. On the reflex mirror, there is a small translucent piece that allows some light to pass through to the sub-mirror. If the image is out of focus, then the sub-mirror splits the light into 2 rays-- the equivalent of two of the same image. These two images are reflected onto the phase detection sensor. The sensor then represents the images and compares them. The goal is for them to be "in phase", or not split-images. Once the phase detection sensor registers a single image (the split-image registered as one), then the subject is in focus, i.e., the farther out of phase the images are, the more out of focus the photo will be. The comparison that the phase detection sensor performs is a computation called a phase difference calculation-- which is carried out by a kind of differential equation. This calculation involves computing the distance from the subject which, at least partially, determines the phase difference as represented in the sensor. Once the sensor has this calculation completed, it sends a correction signal to the lens to make the necessary adjustments. This whole process takes less than a second. The question now is whether the structure and function of PDAF systems in DSLR cameras instantiate the identified properties of control.

[9] For the sake of space, my discussion of autofocus will be limited to phase detection autofocus systems (PDAF). PDAF are probably the most popular systems on the market. For that reason, when one talks about autofocus, they are typically referring to phase detection systems. There are other kinds of autofocus systems available, e.g., contrast detection, and I think that the following arguments for the implementation of control will apply to those as well.

PDAF systems are designed to contribute to the overall production of a focused photo. They contribute to this task by focusing the image such that when the photo is taken, the point of focus is captured appropriately. Therefore, if we assume the action type: *take clear photos*, PDAF systems guide the production of this output by completing a subtask of the action type, namely, focusing the image. In this sense, PDAF systems guide action, but they also reliably produce the appropriate output in contributing to the overall action type.

PDAF systems are particularly well suited for focusing action shots. Taking a clear action shot is no easy feat for a photographer with a manual focus camera. The photographer has to constantly monitor changes in the environment that happen in split seconds. This means that he must flexibility respond to a rapidly changing environment. This burden still exists in the case of DSLR cameras with PDAF systems, but the burden is shifted to the PDAF system instead of the photographer. Because PDAF systems can calculate the phase difference and quickly adjust accordingly, it can keep up with the changes in the environment. Indeed, PDAF systems flexibly, perhaps more so than photographers, adapt to rapidly changing environments to focus the action shot(s).

But PDAF systems also process information about the controllee in focusing an image. As the light ray is reflected off the sub-mirror, it is split into two rays (if the image is unfocused). The PDAF sensor takes in both rays and performs computation over the representations of the images, namely, phase difference calculation. The calculation is used to determine relative lens adjustment for the target output of a focused "through-lens" image. Thus, the correction signal manipulates the lens in virtue of a computation. In responding to a rapidly changing environment, e.g., taking action shots, the PDAF system might continually manipulate the lens to

stay in focus. In being both flexible and manipulative, PDAF systems are able to appropriately guide behavior. In the next section, I argue for some criteria under which control can be said to extend.

4.0 Developing the Theory of Extended Control

In the previous section, I argued that PDAF systems, in virtue of their structure and function, instantiate the properties of control identified on Shepherd's account. Thus making the case that at least some artifacts can implement control. In this section, I hope to glean some criteria from the arguments of the previous sections to develop the minimal benchmarks of extended control into a theory. Here, I am particularly concerned with how control systems form between agent and artifact, *given that the artifact in question is a controller.*

4.1 Information Processing and Autonomous Maintenance of Control

I have emphasized the role of information processing in my account of extended control. It is crucial in explaining how an artifact is a controller. That is, how artifacts can meet the conditions of control set out by Shepherd. But one might reasonably wonder what exactly I have in mind when I invoke information processing. Information is a notoriously slippery concept and without a clear idea of how I am intending to use it, extended control can seem mysterious at best and at worst, a failure. It might be a failure if the right notion of information is not available to me. Take for example the distinction between semantic and non-semantic information. Shannon information is a paradigmatic example of non-semantic information. Shannon information is a probabilistic notion. On this view, information is just the reduction in uncertainty. On the other hand, semantic information is structured and stands for something else. As Piccinini and Scarantino describe it:

> We call 'semantic information' the information a signal carries by reducing uncertainty *about some state of affairs*. In this case, semantic aspects are crucial: what information the signal carries is constitutively related to what the signal stands for (2010, pg.241; emphasis added).

Whereas semantic information involves intentionality, Shannon (and other non-semantic information) do not. Shannon himself notes as much:

> [The] semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is *selected from a set* of possible messages' (1948, p. 379).

Here, the reduction in uncertainty is not about some state of affairs or other, but rather concerns the selection process as a whole, as among the set of possible alternatives.

Semantic information can be further distinguished into two kinds: factual and instructional. Following Fresco (2013), factual semantic information is declarative-- it is about facts. Instructional semantic information, on the other hand, carries directives either imperatively (e.g., do X) or conditionally (e.g., if Y, then X). Notice that in either case (of factual or instructional information), the information that is carried is intentional-- it stands for something else.

For my part then, it seems clear that extended control requires semantic information processing, and not mere Shannon information. To see this, recall my insistence on feedback about the controllee for an artifact's being a controller. Feedback is necessary because it's hard to see how an artifact can satisfy Shepherd's conditions for control without it (see section 2). Shepherd's conditions entail goal-directedness according to at least the approximate goal matching condition. Here, the controller has some goal G (that need not be represented as such) and acts to converge on G. The extent to which it does this is part of the degree of control it

exercises. Feedback about the status of the controllee is central to converging on the goal and flexible repeatability. Taken together, the notions of goal-directedness and feedback entails intentionality. Thus the relevant notion of information processing will be semantic information.[10] If artifacts cannot process semantic information, my account of extended control fails. I intend to show that artifacts, and in particular the PDAF system, process semantic information. To do this, I rely upon causal-indicative semantics (Dretske, 1981; Long, 2014, 2018)[11].

To begin with, note that whatever intentionality an artifact has, it will be derived. We can, following Fred Adams (2010), distinguish between two kinds of derivation: causal and semantic. Causal derivation of content involves causing a system to have a certain symbol-meaning combination. That is to say, the formation of the symbol-meaning combination is causally situated, but the matching between the symbol and meaning is autonomous. On the other hand, when the symbol-meaning combination is not autonomous, it is derived, or preassigned to the system. I assume that the PDAF system's content is both causally and semantically derived by the designers. This is important because if one thought that artifacts could not process semantic information in virtue of the fact that they have derived contents, they would be mistaken because they would be running together causal and semantic derivation.

Having made it plausible that artifacts can process semantic information despite having derived contents, I now turn to the question of what makes an artifact's content semantic as opposed to mere Shannon information. My claim is that the informational content in artifacts are teleologically derived from the design of the artifact. In Fred Adams' terminology, the informational content is both causally and semantically derived by the designers. To say that

[10] Thanks to an anonymous reviewer for pointing this out to me.
[11] I understand that someone might disagree with this theory of semantic content, but I do not have the room to get into alternatives here.

some signal in the system means X, is to say that it naturally indicates X. It was selected to play a role in the system because of this. For the purpose of understanding the system, that is what the signal means.

Causal-indication semantics (CIS) is intended to capture the "stands for" relation of semantic information. Probably the most popular version is Fred Dretske's (1981) indicator semantics. Dretske's project was to reduce semantic content to intentional content by bootstrapping intentionality out of Shannon information. Dretske takes his cue from Grice's natural signs. For example, smoke means fire because smoke is an indicator-- it indicates that there is a fire. Indication is a relation between types of states of affairs; it relates facts (Godfrey-Smith, 1992). Thus Dretske's view is the following:

> The information content of a signal is being expressed in the form "s is F" where the letter s is understood to be an *indexical* or *demonstrative* element referring to some item at the source. What the definition gives us is what philosophers might call the signal's *de re* information content (1981, pg.66).

For Dretske then, a signal carries informational content by indication. But indication cannot occur without causal pathways-- smoked is *caused* by fire. Fire is a Shannon source and the informational content that is carried by the smoke via indication is causally sustained. As Long (2018) puts it:

> [T]here can be no information about a source entity without either information from that entity acquired through physical causal pathways or else by way of causally sustained structural covariance (Long, 2018, pg. 159).

Because information is transmitted via causal pathways only, a source entity is a spatiotemporally structured entity that can causally interact with other physical structures. A

structure naturally indicates something about the cause of its configuration. Thus, semantic content is constituted by indication along those causal pathways from source entities.

Now that we have some tools on the workbench, I want to put them to work on the example of the PDAF system to show that it processes semantic information. In particular, I will focus on the correction signal that the PDAF system generates to focus the lens. Once the system has calculated phase difference by comparing the split images on the image sensors, a correction signal is generated and sent to a lens adjustment mechanism. Here, the signal effectively causes the mechanism to adjust the lens such that the image will be in focus. If the image sensors detect that the images are not in-phase, the process is repeated. Once the subject is in focus (the images on the sensor are registered as in-phase), the system sends a confirmation signal (a beep or green dot in the viewfinder).

Most important for present purposes is the correction signal's causal influence on the lens adjustment mechanism. Firstly, the fact that there is a causal pathway from the image sensors to the lens adjustment mechanism, shows that there is causal indication between source entities. The lens' adjustment (due to the correction signal) indicates something about its cause. So at a minimum, there is semantic content about the cause of the lens' configuration. But we can go further with this. Recall the distinction between factual and instructional semantic information. Factual information is information *about* facts or states of affairs, whereas instructional information carries directives either imperatively or conditionally. Instructional information is used to produce or bring about some state of affair(s). A natural way to interpret the PDAF correction signal then is as instructional information and thereby semantic information. To see this, recall that semantic content is necessarily about something else-- it has intentionality. A

correction signal is *about* the way the lens is to be adjusted. Furthermore, designers selected this function for the PDAF system. Its intentionality is derived both semantically and causally from the selection process.

Because the PDAF system continuously measures for focus, it is a closed-loop system (it receives feedback). Given my arguments above, it follows that the looping entails semantic information processing. Now, we can generalize these insights: for any artifact that is a closed-loop system (has mechanisms for feedback about the controllee selected for by the designers), it will process semantic information. To see this, note that the controller exerts causal influence on the controllee via transmission of instructional semantic information and then receives feedback about the status of the controllee compared to the goal. Indeed, it's hard to understand how a controller could drive the controllee into any of its range of states without instructional semantic information being processed somewhere in the system.

4.2 Expectation

In cases of extended control, the agent expects the artifact to produce the appropriate outputs. In the autofocus case above, the photographer doesn't question the reliability of the autofocus feature. Rather, he expects the feature to perform a certain way. In the same way, an agent expects his motor control to produce the right outputs, say, when he wants to get up from his chair. In the camera case, the photographer expects his autofocus to not fail.

How is it that we come to expect certain outputs from artifacts? I can think of two ways this happens: (i) expectations are culturally or socially nurtured, e.g., marketing, interactions with other people and artifacts, and (ii) our own prior experience with artifacts. Both (i) and (ii) constitute expectations that we have about how artifacts ought to perform when we rely upon

them to carry out some goal-directed behavior. Of course, these expectations are not always accurate. I'm thinking of cases where a person makes a user-error on his computer, but attributes the mistake to the computer and not himself. But even if these expectations don't track reality all the time, this doesn't preclude the artifacts' being a controller. The artifact still has the mechanisms to instantiate control. This does, however, mean that the agent cannot enter into the appropriate delegation relation (more on this in the next section).

Suppose the photographer expects the camera to shoot fire out of the lens. In this case, his overall goal state, perhaps to burn his enemies, will not match up with the camera's function. Thus, the camera cannot control for the subtask of emitting fire. But, importantly, it can control for the subtask of focusing images. It does not lose this function in virtue of an inaccurate expectation. So were the photographer to have the right kind of expectations, the overall goal state would likely match up with the camera's function (and assuming the camera implements this function, the agent's goal-state will match with the behavioral outcome). And the camera, all else being equal, would be a controller for focusing the image. It seems, then, that not just any expectation will do, the expectation needs to be *reasonable* given the artifact's function.

The above discussion illustrates a close connection between expectations of the artifact and the goal-state of the agent. The expectation must be outputs that the artifact can actually produce and that contribute to the goal of the agent. Just as human cognitive control needs to send reasonable motor commands to motor control, agents need to have reasonable expectations for the artifacts they engage with. If cognitive control were to send unreasonable motor commands to the motor control centers, cognitive and motor control would fail to be functionally integrated and thus fail to result in a controlled behavior. Likewise, unreasonable expectations of

an artifact's outputs fails to extend control in a functionally integrated control system where the agent and artifact are parts.

However, artifacts can be *repurposed* for tasks that they were originally not designed to do, and control can still be extended to them. For example, Rodney Brooks invented the Roomba-- a robot that moves around your house and vacuums up messes. There have been many videos made of people repurposing them to fight one another. They attach knives and other objects to the Roomba and put them in an enclosure of some sort. The result is that the Roombas "fight" each other. In this case, the original expectation of the Roomba was to clean up messes. Now, with a few design changes, the expectation is that it will destroy the opponent Roomba in a cage match. The agents' goal-states change too. The agent who wants his floors clean, expects the Roomba to be a controller for the task of cleaning up messes on the floor. The agent who wants to win money in a bet over whose Roomba is superior, expects his Roomba to destroy the opponent. These goal states and expectations make sense given the function of the artifact.

Now sometimes, repurposing an artifact for a different function involves changes to the underlying mechanisms. When this happens, taxonomy-by-mechanism re-enters the picture. Imagine that a DSLR camera with a PDAF subsystem was repurposed for shooting paintballs. It seems clear that major mechanistic changes are necessary to equip the DSLR with this function. But it's not clear whether the camera-paintball-gun is still a controller. The only way to find out is to analyze its behavior in search of control properties and then analyze the underlying mechanisms for what structures realize those properties.

One might object at this point that the expectation condition is too weak. In particular, one might be inclined to think that the agent needs to *know* (rather than merely expect) that the

artifact can implement the desired function. Daniel Dennett (2015) seems to endorse this kind of position in his discussion on an agent controlling a remote-controlled plane:

> Now you can control the direction, height, speed, turning, diving, and banking of the plane. There are many more degrees of freedom, and they are all under your control, but not until you have discovered the parameters of the plane's degrees of freedom and the causal relationships between those parameters and your joystick-moving acts-- an epistemic problem that is often not trivial and must always be solved before control can be affected" (Dennett, 2015, pg.58).

Now, Dennett does not explicitly say that one must *know* the various parameters and relations, but perhaps if one is to successfully execute a barrel roll with this plane, one must know that the plane can do such maneuvers and how to drive it into such a state. There are two questions here: (i) does the agent need to know that the plane can barrel roll? and (ii) does the agent need to know *how* the plane can implement a barrel roll given that it can execute such a maneuver?

Let's consider the second question first. It seems clear to me that the agent needs to know how to drive the plane to execute its barrel roll function. It's not clear to me that the agent needs to know how the plane itself implements that function. To be sure, if the agent desires the plane to barrel roll, he ought to know how to make it do it, e.g., by moving the joystick thus and so. But I do not think that the agent needs to know what those particular joystick movements do to the mechanisms in the plane such that the plane does the barrel roll. A photographer knows how to engage the autofocus feature, but we can imagine that he does not know that the subsystem engages in phase difference calculation. Yet, his images will come out clearly focused.

The first question, however, is a bit trickier: does the agent need to know (and not merely expect) that the artifact in question can implement the function of, say, focusing images when he wants to take clear photos? Knowing, rather than merely expecting the relevant outputs,

presumably is intended to create a strong (if not perfect) match between the distal goal and the overall outcome of the behavior. For if I know that the autofocus system will focus the image when engaged, then it's true that the image will be focused. On Shepherd's theory, we saw that the degree of control one has is proportional to the degree of correspondence between goal and outcome (among other factors). Knowing that the artifact would produce the right outputs for the overall goal-state would, presumably, increase the likelihood of a perfect match. But notice that the artifact only contributes to part of the distal goal, hence knowledge of the outputs would only increase the *likelihood* of a perfect match. Thus, knowing that an artifact will yield the right outputs neither negates nor entails a perfect match between distal goal and overall outcome. Additionally, even if the artifact contributes totally to the distal goal, correspondence between goal and outcome is not the only condition on control. Suppose I expect the autofocus system to focus the image and it comes close, but it is not completely focused. That is, there is an approximate correspondence between distal goal (take clear photos of the runners) and outcome (slightly out of focus of the runners). This does not mean that I am not in control and that the autofocus subsystem is not in control, at least according to Shepherd's account. For the degree of correspondence is not perfect, but it is approximate. And approximate correspondence is not enough to negate (extended) control. I conclude then that knowing that the artifact will implement the desired function, or put otherwise, that the artifact will yield the appropriate outputs, is not necessary on this account.[12]

4.3 Delegation

[12] Although I suppose if one were interested in successful action, or perfect match actions, then knowing would be better than expecting. But again, knowing that the artifact will yield the right output neither negates or entails perfect match. There are other factors to consider. Thanks to an anonymous reviewer for this insightful and challenging objection.

I mentioned in an earlier section that artifacts are dependent upon agents for their contribution to the relevant goal-directed action. Delegation is meant to capture this relation. Thus, delegation is the creation of a causal dependence relation between the agent and the relevant artifact such that when this relation is created, the artifact inherits the (derived) intention to Θ from the agent's (non-derived) intention to ϕ, where ϕ-ing entails Θ.[13] For example, assume a photographer intends to take clear photos of the track runners. His intending this entails focusing the image. We can derive the intention to focus the image from the intention to take clear photos (of the runners). When the photographer interacts with his DSLR camera, he creates a causal dependence relation between himself and the camera-- i.e., were he to not engage with the camera, the camera would not perform its function.

Inheriting a derived intention here means that the artifact's function (a disposition to act in a way specified by the manufacturer) is implemented. In other words, the artifact always has the function to focus the subject, but it does not "intend" to do so until the agent engages with it in a particular way. Tying this back to the information processing section, the designers set up the artifact to process (derived) semantic information in performing its function, but it is the agent who actually gets the artifact to do the processing by engaging with it via delegation. Secondly, because information is transmitted via causal pathways we can say that the delegation relation creates an information transmission channel between the agent and artifact. Indeed, I think this is crucial to how the agent and artifact become a functionally integrated control system. The way the agent interacts with the artifact transmits information to the artifact and

[13] HRI researchers, Miller and Parasuraman (2007), explicitly endorse a delegation model of supervisory control for human-automation systems. As they characterize it: "In short, delegation is a process of assigning specific roles and responsibilities for the subtasks of a parent task for which the delegating agent retains authority (and responsibility)" (pg.64).

back to the agent. For example, what the agent aims at indicates to the camera what is to be focused and the agent's looking through the viewfinder indicates to him whether the subject is focused or not.

Human-robot interaction (HRI) researchers Crandall and Goodrich (2002) provide an interaction scheme for understanding the efficiency of human-robot interaction. This scheme has three parts: autonomy mode, control element, and information element. Autonomy mode refers to the degree of autonomy that the robot has. The control element is used by the human to communicate information to the robot. This is usually a kind of interface that is designed for humans to input functions and commands. Finally, the information element is used by the robot to communicate to the human. Here, the robot is designed to process information through feedback and communicate it back to the human user. It is clear that these researchers suppose an informational link between human and robot for functional integration which promotes efficiency. Relatedly, Bauer et al. (2008) discuss how robots can derive intentions from their human users. Among the many ways, is through manipulative gestures. Manipulative gestures are acts that manipulate objects or motions through which systems can derive a users intention. These gestures are often, though not always, unconscious. Many artifacts have input modalities designed and dedicated for gesture data, e.g., touch screens and camera systems.

I mentioned in the previous section that if an agent has unreasonable expectations given the function of the artifact, then the agent will not enter into the appropriate delegation relation. If an artifact inherits a derived intention to Θ given the agent's intention to $\boldsymbol{\phi}$, then presumably the artifact must be able to perform Θ. Thus, the derived intention corresponds loosely to the agent's reasonable expectations of the artifact's output. Were the agent to have unreasonable

expectations, then the delegation relation would not hold and the agent and artifact would not be functionally integrated. Supposing an agent has an unreasonable expectation, it is very plausible that the way he interacts with the artifact will indicate this. The artifact cannot, at worst, implement its function, and at best, accomplish its goal without the agent's having the right sort of expectations indicated to the artifact.

4.4 The Coupling-Constitution Fallacy

Before closing this section, I want to take the coupling-constitution fallacy head on. Some of what I have said in the delegation section could reasonably be read as committing this fallacy. Before moving on, it would be best to address this worry here.[14]

A critic of extended mental properties might reasonably ask whether I commit the CC fallacy in arguing for extended control. There are two ways a critic might do this. Firstly, the critic might claim (as Adams & Aizawa do for cognition) that one needs a theory of control to make the inference from a causal claim to the constitution conclusion. Secondly, the critic might point out that the delegation relation is a causal dependence relation which confuses a causal relation with a constitution relation. This second way of running the CC fallacy is distinct from the first in the following way. This second way says, "look, even if you provided a theory of control, your delegation relation misses the target because it does not move you to the constitution claim, like you need it to. We need some additional reason for thinking that, for example, the photographer and the camera are a single control system, rather than two causally interacting control systems."[15]

---

[14] I should note that some philosophers have expressed skepticism about the legitimacy of the CC fallacy (Craver, 2007; Kirchhoff, 2014, 2015; Ross and Ladyman, 2010). I am sympathetic to these criticisms, but I set them aside for now. For as I claimed earlier, my plan is to use the objections to extended cognition as constraints on my theory of extended control. Thus, I assume here that the CC fallacy is a legitimate criticism and worth taking seriously.

[15] Special thanks to an anonymous reviewer for pointing out this way of running the objection.

I will now respond to both of these critiques. As for the first, I have provided a theory of control that delivers verdicts about an artifact's status as a controller in section 2. This theory includes three conditions: (i) approximate goal matching, (ii) flexible repeatability, and (iii) no causal deviance. Furthermore, I demonstrated how these conditions are met in a PDAF system by reference to the mechanisms and processes of the system. This then allowed me to say that the PDAF system is a controller. This leads me to the second objection.

On this way of running the objection, while it may be true that the camera and the photographer are controllers, they do not form a unified, singular control system. Rather, they are interacting, but independent controllers. In response, I'm going to draw upon Lynne Rudder Baker's (1999) work on constitution and Robert Wilson's (2005) amendment of her view to accommodate group agency. In slogan form, Baker's account of constitution is that when things with certain properties are in certain circumstances, new things with new properties come into existence. For example, when a piece of marble is subjected to the artist's tools, given a title, and displayed to the artworld, a statue comes into existence. The statue has properties that the marble does not, e.g., it has the property of aesthetically moving people. Baker also thinks that constitution is a contingent relation-- the marble could exist without the statue existing. But when the marble exists, and the right circumstances obtain, the statue exists. Finally, Baker thinks that the marble and the statue are spatially coincident-- they occupy the same regions of space.

Wilson amends this final condition to account for group agents. He has the intuition that, say, when a group of people are in the right circumstances they constitute a corporation. There is a sense in which corporations act, they have a causal impact on the world. But crucially, the

corporation and the group of people are not spatially coincident. They are not spatially coincident because they are not physically bounded. So instead of spatial coincidence, Wilson proposes *agency coincidence.* As he puts it, "two agents are agency coincident at *t* just if they undertake precisely the same actions at *t*" (Wilson 2005, pg.22). So suppose that a corporation is buying out a competitor. The set of individual agents that are involved in making that happen are agency coincident with the corporation. This is because the corporation and that set of agents are engaged in precisely the same action.[16]

One might object here and say that the individuals are all performing different sub-actions. This is right, but Wilson's claim is that the *set* of individuals are engaged in the same action as the group agent, not that each individual is engaged in the same action. The set of individuals is not performing sub-actions, the set can accurately be said to be buying out the competitor, which is the same action that the corporation is performing.

Another objection might be that Wilson has changed the subject-- he's not talking about constitution, he's talking about "schmonstitution". The spatial coincidence condition is a very strong and intimate condition that captures something really important, whereas the agency coincidence condition is weaker and less intimate. In way of response, I want to suggest that perhaps there are many kinds of constitution relations. Wilson (2007) himself, in a different

[16] Wilson states: "To make the case that collective social agents are agency coincident with the collections of individuals that belong to them and that they represent is a large task that I do not propose to undertake here. But I would not raise it as a possibility if I thought that it had no *prima facie* plausibility" (2005, pg.23). I do not want to be interpreted as endorsing the claim that social agents are *actually* agency coincident with the collection of individuals that belong to them. I am borrowing this notion of coincidence to show that the photographer-camera system is agency coincident with the camera and the photographer. I remain agnostic about the truth of Wilson's account.

paper, argues that there are at least two: compositional and ampliative constitution. Wilson defines these as follows:

> When some particular entity *y* is *compositionally constituted* by some entity *x* or some entities the *xs*, *y's* existence is necessitated simply by the state that *x* itself is in or the precise way in which the *xs* are arranged.
>
> By contrast, when some particular entity *y* is *ampliatively constituted* by some particular entity *x* or some entities the *xs*, *y* is an entity whose existence is not necessitated by that of *x* or the *xs*, whatever intrinsic state *x* is in or however the *xs* are arranged (2007, pp.3-4).

An example of compositional constitution, according to Wilson, is the hydrogen and oxygen molecules that constitute water-- water is *nothing more than* the particular arrangements of the molecules. On other hand, the way that a statue is constituted by marble requires more than the mere existence of the marble-- statues *are more than just* the marble. One can now see that we have these two different notions of constitution because of the various relata that we think the constitution relation applies to. It would not be surprising that there are other notions as well. Perhaps the notion that applies in the case of group agency is one of these cases where the relata are such that we have to bring in a different notion of constitution-- one that has agency coincidence instead of spatial coincidence as a condition. After all, it's not incoherent to think that if there are group agents, they would be constituted by a set of individuals.

Having hopefully motivated Wilson's notion of agency coincidence, the question now is whether the other conditions that Baker proposes are met. Here's what Wilson says:

> Consider the collection or group of persons who belong to or are represented by a given collective social agent at a given time. Clearly, that group of persons could exist without that social agent existing. Each member of the board of directors could exist without the social institutions presupposed by the existence of the board itself, or the mechanisms making just those individuals members of the board. But if we have a group of persons, and those conditions are in place—the corporation exists, those persons have been

> appointed to the board, etc.—then we must also have the corresponding collective social agent, in this case, a board of directors. As in the standard cases of constitution, this is guaranteed by the nature of these conditions, and so taking an agent that is a kind of constituent—a group of people—and adding these conditions is metaphysically sufficient to create a collective social agent (2005, pg.23).

If Wilson is right, then group agents can be constituted by the set of individuals that make them up. If a group of individuals that make up an institution are in the right kind of circumstances, are agency coincident with the institution, and the institution could have failed to exist despite the group of individuals existing, then the set of individuals constitute a group agent.

For my part, I intend to show that the photographer-camera system is constituted by the photographer and camera using the conditions that Wilson gives. Firstly, it seems obvious that the system could fail to exist despite the camera and photographer existing. In fact, this kind of system fails to exist often because the kind of circumstances that are required to bring the system into existence (i) sustains the system's existence through time and (ii) are intentionally brought about and dissolved by the photographer. The kind of circumstances I'm referring to here are the conditions that I have been spelling out in this section: the right kind information processing and control maintenance, reasonable expectations, and the delegation relation. Finally, assuming the action description is 'taking focused pictures of a particular subject', the extended photographer-camera system is agency coincident with the photographer and camera. This is because the system and the set of individuals in the system are engaged in the same action. To be sure, I am thinking of agency here in a very general sense. A thing that merely engages in goal-directed behavior can be counted as an agent. So the camera counts as an agent here, capable of acting.[17] If this is convincing, then all of the conditions that Wilson proposes have

[17] Wilson has a similar, but more radical view: "I take agents to be individual entities that are capable of acting in the world, and that they typically do so act. They are differential loci of actions. I am happy to be quite pluralistic about

been met and we are licensed to say that the photographer and camera constitute a unified control system.

In this section, I have been teasing out some criteria that I think are crucial to the notion of extended control. These criteria are by no means necessary and sufficient conditions, but they go some way towards fleshing out an account that corresponds with our intuitions from the autofocus case. Just to recapitulate those criteria: (i) information processing (about a controllee) and autonomous control maintenance, (ii) expectation, and (iii) delegation.

5.0 Extending Extended Control to Cognition and Group Agency

Before ending this paper, I want to briefly consider a consequence for extended cognition that might fall out of my theory of extended control. This consideration will merely gesture towards a way that one can replace the question of whether cognition extends with the question of whether control has extended.

Extended cognition theorists think of themselves as offering a novel and superior way of explaining agents' complex interactions with artifacts by thinking of the agent and artifact as a cognitive system. Among the many objections to extended cognition, one is that these theorists have not provided a theory of cognition to deliver verdicts on whether it has indeed extended. This is a kind of epistemic problem-- we just do not know enough about cognition to say whether it extends or not (although Adams & Aizawa think we do know enough about it, and it's true that it does not extend). Secondly, there is a logical problem: the CC fallacy. One cannot directly infer a constitution claim from a causal claim without supplying the missing premise from a

---

the kinds of agent there are in the world. There are *physical* agents, including elementary particles and atomic elements, everyday physical objects, such as tables and rocks, and larger and more distant objects, such as stars and tectonic plates" (2005, pg.11).

theory of whatever property is claimed to extend. Another objection seems to me to be that it is just counterintuitive that certain artifacts are cognitive when one interacts with them, e.g., are Otto and his notebook *both* cognitive?. This seems to be an ontological problem-- it's just strange that notebooks would be cognitive. Indeed, Adams & Aizawa (2001, 2010) claim they are defending common sense in arguing against the extended cognition hypothesis (see especially 2001, pp. 44-46).

Extended control, however, avoids these objections and still makes good sense of agents' interactions with artifacts. It avoids the epistemic problem because control is much better understood than cognition is and, in this work in particular, I offer a theory of control. By supplying a theory of control, I also avoid the logical problem. Shepherd's theory makes predictions about what sorts of things can be controllers and guides mechanistic analysis to deliver verdicts about an artifact's status as a controller or not. Extended control avoids the ontological problem because it is *not* counterintuitive that artifacts be controllers. As I have noted before, it is commonplace in engineering to think of artifacts as controllers as well as in ordinary language, e.g., "the thermostat controls the temperature".

Extended cognition theorists are sometimes motivated by the notion of "cognitive relief" or "offloading". When a human interacts with an artifact, the artifact usually takes over some of the cognitive workload that the human would have had were he to not interact with the artifact. This has led some theorists to conclude that the artifact is engaged in a cognitive process, so cognition extends. HRI researchers Crandall and Goodrich (2002) discuss the human cognitive load parameter in connection with the level of autonomy that a robot has. The idea is that the more cognitive relief afforded to the human by the robot, the higher the level of autonomy. But

of course, the higher the level of autonomy, the higher the level of control the robot has. This is because the artifact is less dependent on the human in executing its functions. Which means that it will have to be more flexible and efficient. So we can talk about cognitive relief in humans by noting the level of autonomy that the system has instead of committing to extended cognition. And when one shifts to talk of levels of autonomy, they are implicitly committed to control being extended.

In sum, I think the consequence I try to develop here is more about theoretical virtues. We ought to prefer extended control insofar as it can capture the explanatory goals of extended cognition theories while avoiding the epistemic, logical, and ontological problems. In other words, extended control provides a language for discussing complex agent-artifact interactions without getting stuck on the thorny metaphysical issues associated with extended cognition.

Another interesting question might be the following: are there limits to how much of a plan an artifact can implement for an agent? In other words, what might be the upper bounds of extended control? Consider a complex artifact that has the ability to modify or override the derived intention that it inherits in virtue of the delegation relation. Humans have this ability--imagine a boss that delegates a task to an employee, but the employee modifies the plan (or even the task) to better satisfy some distal goal. It seems in principle that an artifact could do this too without challenging extended control. All that would be required is that the agent expect the artifact to implement this function of overriding or modifying the derived intention it inherits. So suppose my GPS and I enter into an extended control system and I expect it to correctly yield the fastest route to my destination. Some GPS systems are programmed to supply the fastest route and it modifies the route in accordance with this program. As long as I expect that it will modify

the derived intention in some way, there does not seem to be a breach of extension. Things become problematic if I do not expect it to modify the plan, or if I do expect it to do so, and it radically strays from the plan such that my distal intention is not even approximately matched. In the former case, it seems that my GPS and I are not functionally integrated. If I did not expect my motor control to override my action plan given by cognitive control, then I certainly would not be in control of the resultant action. In the latter case, there is no longer a correspondence between goal and outcome.

When groups of people are considered group agents, it makes sense that they'd expect each other to modify or override previous plans-- this might be a part of the deliberation process. In fact, in some cases, particular members of the group are assigned this function: to modify or override plans on behalf of the group, e.g., supervisors. Group agents sometimes negotiate and coordinate what they ought to do to achieve their goal and this is unsurprising, i.e., they expect it. Similarly, in extended control systems with complex artifacts, it will also be unsurprising if artifacts enter into this coordinating relationship with agents so long as the agent expects it to do so. The literature on group agency and collective intentionality (for review see Gallotti and Huebner, 2017; Huebner, 2014) sometimes is cashed out in terms of extended cognition-- that is, if extended cognition is true, then group agency is possible. Given the above arguments (that we should prefer extended control over extended cognition), it might be a fruitful question whether extended control can serve as a better basis for developing a theory of group agency.

6.0 Conclusion

In this paper, I have argued for a theory of extended control to explain our interactions with artifacts in goal-directed behaviors especially when both the agent and the artifact

implement controlled processes. I gave conditions under which control could be implemented in artifacts by taxonomy-by-mechanism. At bottom, for an artifact to be a controller, is for it to process information about the controllee to appropriately guide the controllee into the desired state. This information processing explains the artifact's capacity to be successful across a wide range of counterfactual situations, i.e., flexible repeatability. I also explained how an agent and artifactual controller can be functionally integrated in an extended control system. The agent must have reasonable expectations about the artifact's output. Secondly, the agent must delegate control to the artifact by creating a causal dependence relation that involves the artifact inheriting a derived intention entailed by the agent's overall intention. Finally, I considered the upper bounds of extended control and how this theory ought to be preferred to extended cognition.

References

Adams, F. (2010). Information and knowledge à la Floridi. *Metaphilosophy*, *41*(3), 331-344.

Adams, F. and Aizawa, K. (2008). *The bounds of cognition*. John Wiley & Sons.

Adams, F., & Aizawa, K. (2010). Defending the bounds of cognition. *The extended mind*, 67-80.

Baker, L. R. (1997). Why constitution is not identity. *The Journal of Philosophy*, *94*(12), 599-621.

Baker, L. R. (1999). Unity without identity: A new look at material constitution. *Midwest Studies in Philosophy*, *23*, 144-165.

Bauer, A., Wollherr, D., & Buss, M. (2008). Human–robot collaboration: a survey. *International Journal of Humanoid Robotics*, *5*(01), 47-66.

Bermúdez, J.P. (2017). Do we reflect while performing skillful actions? Automaticity, control, and the perils of distraction. *Philosophical Psychology*, *30*(7), pp.896-924.

Buskell, A. (2015). How to be skillful: opportunistic robustness and normative sensitivity. *Synthese*, *192*(5), pp.1445-1466.

Christensen, W., Sutton, J. and McIlwain, D.J. (2016). Cognition in skilled action: Meshed control and the varieties of skill experience. *Mind & Language*, *31*(1), pp.37-66.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7-19.

Crandall, J. W., & Goodrich, M. A. (2002). Characterizing efficiency of human-robot interaction: A case study of shared-control teleoperation. In *IEEE/RSJ international conference on intelligent robots and systems* (Vol. 2, pp. 1290-1295). IEEE.

Craver, C.F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of science*, *68*(1), Pp.53-74.

Craver, C. (2007). Constitutive explanatory relevance. *Journal of Philosophical Research*, *32*, 3-20.

Dennet, D. (2015). *Elbow room: The varieties of free will worth wanting*. mit Press.

Dow, J.M. (2017). Just doing what I do: on the awareness of fluent agency. *Phenomenology and the Cognitive Sciences*, *16*(1), pp.155-177.

Dretske, F. (1981). *Knowledge and the flow of information.* Cambridge, Ma: MIT Press.

Dretske, F. (1994). If you can't make one, you don't know how it works. *Midwest studies in philosophy*, *19*(1), 468-482.

Fischer, J.M. (1999). Recent work on moral responsibility. *Ethics*, *110*(1), pp.93-139.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral Responsibility*. Cambridge University Press.

Fong, T., Thorpe, C., & Baur, C. (2001). *Collaborative control: A robot-centric model for vehicle*

teleoperation (Vol. 1). Pittsburgh: Carnegie Mellon University, The Robotics Institute.
Fresco, N. (2013). Information processing as an account of concrete digital computation. *Philosophy & Technology*, *26*(1), 31-60.
Fridland, E. (2014). They've lost control: Reflections on skill. *Synthese*, *191*(12), 2729-2750.
Fridland, E. (2017). Skill and motor control: intelligence all the way down. *Philosophical Studies*, *174*(6), 1539-1560.
Fridland, E. (2019). Intention at the Interface. *Review of Philosophy and Psychology*, 1-25.
Gallotti, M. and Huebner, B. (2017). Collective intentionality and socially extended minds. *Philosophical Psychology*, *30*(3), pp.251-268.
Godfrey-Smith, P. (1992). Indication and adaptation. *Synthese*, *92*(2), 283-312.
Grush, R. (1997). The architecture of representation. *Philosophical Psychology*, *10*(1), pp.5-23.
Haugeland, J. (1998). Mind embodied and embedded. In Haugeland, J. (ed.), Having Thought. Cambridge, MA: Harvard University Press, pp. 207–37.
Huebner, B. (2014). *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford University Press.
Hutchins, E. (1995). *Cognition in the Wild*. MIT press.
Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*(1), 43-69.
Kamppinen, M. (1988). Intentionality and information from an ontological point of view. *Philosophia*, *18*(1), 107-118.
Kirchhoff, M. (2014). Extended cognition & constitution: Re-evaluating the constitutive claim of extended cognition. *Philosophical Psychology*, *27*(2), 258-283.
Kirchhoff, M. D. (2015). Extended Cognition & the Causal‐Constitutive Fallacy: In Search for a Diachronic and Dynamical Conception of Constitution. *Philosophy and Phenomenological Research*, *90*(2), 320-360.
Long, B. R. (2014). Information is intrinsically semantic but alethically neutral. *Synthese*, *191*(14), 3447-3467.
Long, B.R. (2018). A Scientific Metaphysical Naturalisation of Information. PhD Thesis. University of Sydney.
Menary, R. (2006). Attacking the bounds of cognition. *Philosophical Psychology*, *19*(3), 329-344.
Menary, R. ed. (2010). *The extended mind*. Mit Press.
Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human factors*, *49*(1), 57-75.
Newen, A., De Bruin, L., & Gallagher, S. (Eds.). (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.
Noë, A. (2004). *Action in perception*. MIT press.
Papineau, D. (2015). Choking and the yips. *Phenomenology and the Cognitive Sciences*, *14*(2), pp.295-308.

Pavese, C. (2016). Skill in epistemology II: Skill and know how. *Philosophy Compass*, *11*(11), pp.650-660.

Pavese, C. (2017). Know-how, action, and luck. *Synthese*, pp.1-23.

Piccinini, G. and Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), pp.283-311.

Piccinini, G., and Scarantino, A. (2010). Computation vs. information processing: why their difference matters to cognitive science. *Studies in History and Philosophy of Science Part A*, *41*(3), 237-246.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of biological physics*, *37*(1), 1-38.

Ross, D., & Ladyman, J. (2010). The alleged coupling-constitution fallacy and the mature sciences. *The extended mind*, *155*, 166.

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, *22*(1), 1-19.

Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. Mit Press.

Rupert, R.D. (2004). Challenges to the hypothesis of extended cognition. *The Journal of philosophy*, *101*(8), pp.389-428.

Rupert, R.D. (2009). *Cognitive systems and the extended mind*. Oxford University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shepherd, J. (2014). The contours of control. *Philosophical Studies*, *170*(3), pp.395-411.

Shepherd, J. (2015a). Conscious control over action. *Mind & language*, *30*(3), pp.320-344.

Shepherd, J. (2015b). Deciding as intentional action: Control over decisions. *Australasian journal of philosophy*, *93*(2), pp.335-351.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. MIT press.

Sheridan, T. B. (2011). Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *41*(4), 662-667.

Silberstein, M., & Chemero, A. (2012). Complexity and extended phenomenological‐cognitive systems. *Topics in cognitive science*, *4*(1), 35-50.

Szafir, D., Mutlu, B., & Fong, T. (2017). Designing planning and control interfaces to support user collaboration with flying robots. *The International Journal of Robotics Research*, *36*(5-7), 514-542.

Wilson, R. A. (2005). Persons, social agency, and constitution. *Social Philosophy and Policy*, *22*(2), 49-69.

Wilson, R. A. (2007). A Puzzle About Material Constitution and How to Solve It: Enriching Constitution Views in Metaphysics. *Philosopher's Imprint*, *7*(5).

Wu, W. (2016). Experts and deviants: The story of agentive control. *Philosophy and Phenomenological Research*, *93*(1), pp.101-126.