The Extended Cuckoo

The Extended Mind at 25 in conversation with The Extended Phenotype at 41

David Spurrett (UKZN)
spurrett@ukzn.ac.za

Abstract

Arguments that cognition or minds can be extended regularly invoke an analogy with Dawkins' argument that phenotypes can be extended. I argue that there are two neglected ways in which those two boundary-breaking theses are complementary. Much of the argument of *The Extended Phenotype* concerns parasite phenotypes expressed in the behaviour of host organisms. But the options Dawkins considers for this extended manipulation are cognitively internalist. If we view cognition as extended, we can recognise a wider range of vulnerabilities for exploitation. On the other hand, the analogies drawn with Dawkins almost always emphasise the benefit to the individual agent in being cognitively extended. Taking Dawkins' concerns about manipulation and exploitation more seriously leads to a more contested, less optimistic picture of extended cognition and minds. This second line of thinking follows Sterelny's lead, but I argue that hostility presents worse and more pervasive problems that he allows.

Keywords: Extended Mind, Extended Phenotype, Hostility, Scaffolding

1. Introduction

Similarities between the claims that cognition or minds can be extended, and that phenotypes can, have regularly been noted. Clark, for example, has said that the argument for extended cognition is *analogous* to the case for extended phenotypes, and that viewing cognition as extended is "taking an 'extended phenotype' view of the mind" (Clark, 1998). When claiming there is an analogy Clark often cites Dawkins (1982) and refers to his example of the spider's web. In *Supersizing the Mind*, he says that just as "it is the spider body that spins and maintains the web that then (following Dawkins 1982) constitutes part of its own extended phenotype, so it is the biological human organism that spins, selects, or maintains the webs of cognitive scaffolding that participate in the extended machinery of its own thought and reason." (Clark, 2008:123). He makes similar references to an analogy and the spider in *Being There* (1997:218,246), and in other papers (Clark, 2007:176; Clark, 2012). Others who draw the parallel include Menary (2010:14) and Shapiro who says that philosophers arguing for extended minds "are, in essence, applying Richard Dawkins's idea of an extended phenotype to the individuation of minds" (Shapiro, 2004:219).

These are curiously partial and optimistic responses to Dawkins, underplaying antagonism and conflict to focus on benefit to the individual who saves time or effort, improves accuracy, or otherwise gains cognitively. Much of *The Extended Phenotype*, after all, focuses on manipulation of host behaviour by parasites, including cuckoos. How important this is to Dawkins can be seen in his remark that parasite influence on host behaviour is his "personal epitome of Darwinian adaptation, the *ne plus ultra* of natural selection in all its merciless glory" (Dawkins, 2012:xi). If arguing for extended cognition or minds is applying a 'parallel form of reasoning' to the case for extended phenotypes, it is striking that defences of extended cognition are so often presented as undiluted good news. This is not merely the epistemic good news that we'll finally see ourselves aright, but the good news regarding the plethora of cool tools, gadgets and toys that make our cognitive lives easier and more reliable than they would otherwise be, and carry much of the explanatory burden for human distinctiveness (e.g. Clark, 1997; Clark 2003). Arguments that cognition can be extended, embodied, and so forth also typically focus on the beneficial contribution external factors can make *to the individual subject,* or sometimes to the co-operating group (Hutchins, 1995). Aagaard goes so far as to maintain that thinking about 4E cognition involves a 'dogma of harmony' in which "all entities are presumed to cooperate and collaborate" (2020:1; see also Slaby 2016 on the 'user/resource model').

I'm not persuaded that there's a dogma at work here, though there is a tendency to accentuate the positive. The most optimistic defenders of extended cognition occasionally make time to consider possible downsides, including Clark devoting a chapter of *Natural Born Cyborgs* to the topic of 'Bad Borgs' (2003, Chapter 7), and making occasional references to the possibility that scaffolding might not *always* be

beneficial (e.g. Clark, 2002:29; Clark, 2010:58f). We don't need to suppose a dogma to make sense of the optimistic pattern, because a natural way to argue that some cognitive phenomenon can be extended is to take a compelling example of someone in some psychological state, then show how that state could be embodied, situated, extended, etc. The famous Otto and Inga example in Clark and Chalmer's (1998) illustrates this pattern. The optimistic effect, then, could be a product of dialectical strategy. It remains striking that the optimism appears undimmed when defenders of extended cognition draw connections with *The Extended Phenotype*, given the crucial importance of manipulation and conflict to Dawkins' position. Here I explore two neglected ways in which extended phenotype thinking and extended cognition are complementary. Extended cognition can give extended phenotype thinking an expanded set of routes to manipulation, and in return receive a robust argument against optimism. The position I defend is not that the spider in its web *isn't* one iconic instance of the extended phenotype, and an analogy for the extended mind. It is both. But the parasitic cuckoo in the reed-warbler's nest is if anything a more iconic case of an extended phenotype, one that is given far more attention by Dawkins[1] and that suggests different and more antagonistic analogies for extended cognition and minds.

Here's the plan. In section (2) below I review Dawkins' argument for extended phenotypes, particularly his treatment of host manipulation by parasites. In section (3) I identify two ways in which cognition can be extended which form the focus of what follows. In section (4) I argue, first, that many instances of host manipulation by parasites can and should be recognised as other-directed *hostile epistemic actions* ('hostile' in the sense of Sterelny, 2003). That paves the way for arguing, second, that genuinely cognitively distributed agents are vulnerable in ways that Dawkins didn't imagine, through manipulation of their external cognitive processes. In section (5) I briefly show how these considerations about hostility, including extended hostility, look when freed of any connection with gene-centrism. They look bad, and I argue that they will tend to be as bad as they *can* be.

2. Extended Phenotypes

The extended phenotype thesis develops the gene-centric perspective on natural selection, which prioritises the interests of replicators (in being replicated) over the interests of the organism in which they occur (Dawkins, 1976a). The gene-centric perspective is defended against individual-level selectionism partly by theoretical considerations, such as that only relatively digital gene sequences have the right level of copying fidelity and potential 'immortality' to be the target of selection, and by its capacity to explain cases where genes replicate at the expense of the interests of the host organism. The 'Medea' gene in flour beetles, for example, causes developing young of heterozygous parents to die during development, favouring the gene but very much not serving the parents (Beeman et al., 1992). Despite such cases, the

[1] Cuckoos are mentioned on over five times as many pages of *The Extended Phenotype* as spider webs, and parasites – which are the primary topic of two chapters – on over ten times as many (Dawkins 1982).

gene-centric and individual organism viewpoints often approximately agree when the interests of the replicators in being replicated coincide with those of the whole organism in reproducing. None of what follows depends on *endorsing* gene-centrism.

The phenotype of an organism is commonly understood as the *bodily* expression, in morphology, physiology and behaviour, of the genotype, and traditionally thought of as bounded by the skin. Dawkins argues that this is too restrictive. Some organisms reliably modify the world beyond their bodies in ways that matter to their biological success, such as beavers building dams and – yes – spiders building webs. Their capacities and dispositions to do this can have as genetic a basis as features of their bodies and dispositions. When they *do*, Dawkins argues that they should be considered part of the phenotype irrespective of whether they are inside the body of the individual carrying the gene. Dawkins' rather liberal notion of a 'gene for' something indicates a replicator that under typical conditions increases the likelihood of some outcome, and where typical includes the genetic environment, developmental processes and external context (Dawkins, 1982).

Dawkins classifies extended phenotypes into three categories. Beaver dams and spider webs illustrate the first: constructions or artefacts. The second and third involve manipulating the behaviour of other organisms (see also Krebs & Dawkins, 1984). The key idea here is that behaviour by one individual might serve replicators in another. Dawkins discusses a variety of real and hypothetical examples, including simple ones such as the increased tendency of those infected by rabies to bite, which would plausibly serve transmission of the pathogen, carried in saliva, but doesn't benefit the host (Dawkins 1982:220). He offers as the 'central theorem' of his view that: "An animal's behaviour tends to maximize the survival of the genes 'for' that behaviour, whether or not those genes happen to be in the body of the particular animal performing it." (Dawkins, 1982:233)

Some parasites live inside their hosts, and Dawkins maintains that any nervous system "can be subverted if treated in the right way" (1982:69) and "is vulnerable to manipulation by a clever-enough pharmacologist" (1982:71). Dawkins also knows that some parasites operate from outside hosts, including cuckoo brood parasites. Here he argues that "there is no sensible reason why cuckoo genes should not be said to have phenotypic expression in a reed warbler's body" (1982:227). The fact that "the cuckoo does not live inside the reed warbler's body" doesn't *prevent* influence merely transforms the routes it must take. The cuckoo chick "has to rely on other media for its manipulation, for instance sound waves and light waves." This leads to his third category of extended phenotype, and the second for manipulative behaviour control: genetic 'action at a distance'. A parasitised reed-warbler parent is, he says, "an active, complex machine, with sense organs, muscles and a brain". Consequently, the cuckoo chick has to do more than "have its body inserted in the host's nest", it also has to "infiltrate the defences of the host's nervous system, and its ports of entry are the host's sense organs." (Dawkins, 1982:68)

Cuckoo brood parasitism evolved independently at least three times (Payne, 2005) and the details of the various arms races between host and parasite vary regionally. In some egg mimicry is especially highly developed, suggesting investment by hosts in discriminating cuckoo eggs. As Dawkins notes, the challenge of chick discrimination is harder. Cuckoo chicks eject other eggs and hatchlings, impeding comparison or application of a 'reject the odd one out' rule. They also present exaggerated versions of feeding cues. Rufus bush chats in Spain bring more food to chicks of their own whose mouths have been dyed orange as cuckoo chicks there are, but orange dye makes no difference to reed-warbler parent birds in England. Some cuckoo chick vocalisations are more frequent than those of the birds they displace, perhaps sounding like a nest with several chicks in it, and reed-warblers in England bring more food in response to recordings of cuckoo chick begging calls no matter what is in their nest (Alvarez, 2004; Davies, Kilner & Noble, 1998; Davies, 2015, Ch. 10). As I suggested above, the cuckoo chick exploiting its host by manipulating behaviour over a distance is as much an emblem of the extended phenotype as the web-spinning spider, perhaps more so. This isn't surprising, since of one of Dawkins' most central claims is that adopting a gene-centric approach will bring a wider variety of conflictual and manipulative relationships into focus, including ones where the interests of individual organisms are subverted.

A striking, although not frequently noted,[2] feature of Dawkins' arguments is their functionalist character. I don't mean to say that Dawkins' endorses some specific formulation of functionalism as found in the philosophy of mind or science.[3] But the general functionalist commitment is that some things are individuated by their causal or functional roles, accompanied by the anti-reductionist thought that the role-filling can be done in different ways, i.e., that functions can be *multiply realised*. Michael Wheeler calls the notion of multiple realisability a "chauvinism-busting property" (2010: 248), in allowing robots and aliens with different biological makeup to share psychological properties with ourselves, and argues that extended functionalism is a functionalism devoid of prejudices about where the realisers might be found, favouring in instead a "*locationally uncommitted account"* of the implementation of some function (2010:253).

Thus understood, Dawkins clearly makes broadly extended functionalist arguments about phenotypes, appealing to causal role as a principle of individuation that trumps hunches about the boundary of the skin, and arguing that phenotypic effects can be multiply realised. In illustration of the first point, when contrasting the 'conventional' and extended geneticist he says that both make a relatively arbitrary decision about where to draw the line between embryology and the expression of the phenotype, but that the

[2] Ågren's thorough (2021) exegesis of Dawkins' gene centrism, for example, doesn't contain the words 'functionalism' or 'functionalist'.

[3] Dawkins engaged directly with functionalist thinking in an earlier piece on hierarchy and control. Citing Simon (1973), Dawkins says that "computers with the same programming instruction set are in an important sense isomorphic in principle, even though their wiring diagrams may be utterly different, one employing valves, another transistors and the third integrated circuits; how all three work is best explained without reference to particular hardware at all" (Dawkins 1976b:7). Behavioural ecologists, he reasons, "need *software explanations* of behaviour." (1976b:8). Earlier work by Simon (e.g. Newell, Shaw and Simon, 1958) anticipates philosophical functionalism (see Rupert 2016).

conventional one "makes the further arbitrary decision to cut off all chains at the point where they reach the outer wall of the body" (1982:231, 237-8). This is extended functionalism about phenotypes. On the second point, Dawkins appeals to multiple realisation, saying that "an animal artefact, like any other phenotypic product whose variation is influenced by a gene, can be regarded as a phenotypic tool by which that gene could potentially lever itself into the next generation. A gene may so lever itself by adorning the tail of a male bird of paradise with a sexually attractive blue feather, or by causing a male bower bird to paint his bower with pigment crushed in his bill out of blue berries. The details may be different in the two cases but the effect, from the gene's point of view, is the same" (1982:199).

I've said that the means that Dawkins considers for extended phenotype influence on behaviour, in his original (1982) treatment, and in later discussions (e.g. Dawkins 2004) are striking in being cognitively internalist. In support of that claim, notice that Dawkins only imagines that the downstream effects of genes influencing the behaviour of other organisms enter through the front door of the sense organs or via an internal service elevator acting on the brain, like a drug or short circuit. Dawkins' own distinction between two kinds of genetic influence on the behaviour of other organisms corresponds to these two internalist options: Influence is either 'action at a distance' via sensory cues, or by parasites within the body of the host.[4] In 1982 that might have seemed to survey the available options. Not anymore, which brings us back to extended cognition.

3. Extended Cognition

Defenders of extended cognition claim that "[c]ognitive processes ain't (all) in the head!" (Clark & Chalmers 1998:8). Part of Clark & Chalmers' argumentative strategy is to emphasise what they call "the general tendency of human reasoners to lean heavily on environmental supports" (1998:8). Two key versions of this reliance will be important in what follows: epistemic actions, and external processing.

Kirsh distinguishes *epistemic actions*, the function of which is to change a cognitive state, from pragmatic actions which advance towards practical goals (Kirsh, 1995; Kirsh and Maglio, 1994). Manually rotating a jig-saw puzzle piece while moving it around a partially completed puzzle, thereby replacing challenging mental rotation with simple visual comparison to determine fit, is an exemplary epistemic action. Kirsh provides other illustrations, including a cook organising partly prepared ingredients in a workspace to reduce search time when assembling a dish by simplifying perception, or reducing measurement and planning when assembling a dish. Such actions aren't the *conclusion* of completed cognitive processes, but part of how cognition gets done. Manipulation of external objects can replace or

[4] The persistence of internalism in Dawkins' thinking is exemplified in his 'pipedream' about an 'Extended Phenotypics Institute' having three wings: "the Zoological Artefact Museum (ZAM), the laboratory of Parasite Extended Genetics (PEG), and the Centre for Action at a Distance (CAD)" (2004:394).

support operations with internal representations, reduce cognitive burdens, improve accuracy, etc. One can accept this without committing to *general* anti-representationalism about inner cognition.

Hutchins (1995) emphasises how tools and external media, including slide rules and navigation charts, allow significant processing to be conducted outside the head. Hutchins describes many examples of artefacts which transform the computational demands of navigation, including the 'three scale nomogram' which converts multiplication and division operations relating distance, time, and speed into drawing a straight line connecting two known values and crossing all three scales (1995:148). Slide rules on which several logarithmic scales — some mobile in relation to others — are marked allow various computations to be performed by manipulating the scales and are "useful precisely because the cognitive processes required to manipulate them are not the computational processes accomplished by their manipulation. The tasks facing the tool user are in the domain of scale-alignment operations, but the computations achieved are in the domain of mathematics" (1995:170-1). Navigation maps distort some aspects of physical space so that straight lines on the map pick out locations with shared directional relationships at the cost of accurately representing scale (1995, Chapter 3).

Defenders of extended cognition, which comes in many forms, share a picture of cognition as looping out into the environment, and of action itself, as well as tools and external media, as sometimes significantly cognitive. The two ideas I need to take from this, which by no means exhaust the territory of extended cognition, are, first, that of epistemic actions coupled with selective anti-representationalism, and second the idea that external media can sometimes carry a significant processing load. Proponents of extended cognition needn't insist on narrow parity between how inner and outer processes work but can accommodate a variety of relations of coupling and complementarity, as well as extended processes dependant on training and norms (Sutton, 2010; Menary, 2010; Clark, 2008).

In section (2) I noted the functionalist character of Dawkins' arguments for extended phenotypes. Defences of extended cognition are often functionalist. Clark and Chalmers reason that if "as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process" (1998:8). Menary (2010) later dubbed the form of reasoning here the 'Parity Principle', and discussion of it involves some complications about specific versions of functionalism that needn't detain us. Clark later argued there are two different functionalist moves here. One is a "commonsense functionalism" about coarse-grained folk categories, the other a fine-grained account of the "actual flow of processing and representation in the (possibly extended) physical array that *realizes* the course functional role" (Clark, 2008:88-89, citing Braddon-Mitchell and Jackson, 2007). What matters here is commitment to some kind of extended role-functionalism, shared with the case for extended phenotypes, amounting to open-mindedness about how and where some function might be realised.

Earlier I emphasised Dawkins' delight in and emphasis on conflictual and manipulative cases, and his regarding parasite influence on host behaviour as his "personal epitome of Darwinian adaptation, the *ne plus ultra* of natural selection in all its merciless glory" (Dawkins 2012:xi). As we saw, this motivates the 'central theorem' of extended phenotype thinking, that an "animal's behaviour tends to maximize the survival of the genes 'for' that behaviour, whether or not those genes happen to be in the body of the particular animal performing it" (Dawkins 1982:233). Dawkins retained cognitive internalist assumptions when illustrating his central theorem. Next, I explore the consequences of being extended functionalist about *both* cognition and manipulation.

4. The Distributed Cuckoo

4.1 Hostile Epistemic Actions

Reviewing Clark's *Being There* (1997), Sterelny notes the analogy between extended cognition and extended phenotype. Unlike the references to an analogy reviewed at the start of this paper, Sterelny includes parasite manipulation of behaviour in his explanation of extended phenotypes. He observes that approaches associated with Brooks (1991), and partially endorsed by Clark, that the world can be used as 'its own best model' rather than relying on internal representations, "treat the world as benign, or at worst indifferent" (2000:210). He argues, though, that the world, containing other agents, including many with competing and conflicting interests, who stand to gain from deception and confusion, is "frequently hostile" (2000:211). Sterelny concludes that any model of cognition "that leaves out hostility is not a model of cognition in the wild" (2000:215).

Hostility is Sterelny's term of art, differentiating informational environments. In an *informationally transparent* environment, signals an organism can detect are reliably good occasions for behaviours it can produce, so cue-driven behaviour will be successful (Sterelny, 2003:20). Environments aren't, however, reliably transparent. When relevant features of the environment "map in complex, one to many ways onto the cues [an organism] can detect" they are *informationally translucent* (Sterelny, 2003:21). Sometimes translucency is not the result of indifferent heterogeneity but is produced by other living things with competing interests, in which case he says that the environment is *informationally hostile*. Hostility is the expression of antagonistic or competing interests interests through information.

Sterelny regularly credits Dawkins (1982) when introducing hostility, and the notion is unsurprisingly able to accommodate Dawkins' examples of parasitic control through cue presentation, which all concern some agents in the world manipulating and exploiting others through how they present and behave. The cuckoo is a shared example, discussed by both (e.g. Sterelny 2003, Chapter 2; Dawkins 1982 p. 68f). Sterelny's argument from hostility doesn't require gene-centrism because hostility isn't articulated as a

gene-centric notion. It is effectively agnostic about what the competing interests are, even if most of Sterelny's examples focus on interactions between individual organisms.

Part one of *Thought in a Hostile World* is a critique of the view that efficient cognisers can do without representations by relying on the world to be itself, and to cue behaviour appropriately. Sterelny's treatment focuses on abstract design features of cognitive systems. He doesn't *deny* that there can be gains from cue-bound control, but emphasises that those gains aren't guaranteed, that being cue-bound is a source of vulnerability when the cues can be faked or manipulated, and hence that representational cognition can plausibly evolve and pay its way because stored information is less vulnerable to deceptive cues. The obvious benefit of detection systems that yoke behaviour to detection of a single cue is that they're simple and cheap to build and run. Their great weakness is that other agents manufacturing the cue can elicit the behaviour. Female *Photuris* fireflies, for example, produce the mating signals of females of other firefly species to attract, kill and eat males of those species (Sterelny, 2003:15). Ants using the *absence* of pheromones to distinguish (and attack) invaders are exploited by parasitic beetles producing or mimicking the pheromones and get fed by performing food-eliciting gestures (Sterelny 2003, p. 15). And the emblematic cuckoo drives reed-warbler hosts to respond to exaggerated versions of the appearance and behaviour of their own chicks. For male fireflies of a non-*Photuris* species, for ants invaded by pheromone-secreting and gesticulating beetles, and for the reed-warbler working to feed a cuckoo, the world very much *isn't* its own best representation.

We can and should regard many cases of manipulation by means of hostile cue manufacture as epistemic actions. Kirsh and Maglio introduce epistemic actions by saying that they are "used to change the world in order to simplify the problem-solving task" (1994:513). Soon after that they gloss them as "physical actions that make mental computation easier, faster, or more reliable" and specify that they are "*external actions* that an agent performs to change his or her own computational state" (1994:513-514). While the first more compressed formulation is agnostic about whose cognitive demands are changed, the second demands that actions are epistemic, and beneficially so, *for the agent performing them*. If we relax that restriction, and allow that epistemic actions *needn't be self-directed*, we can accommodate many cases of hostility, including some of Dawkins' own illustrations of parasite manipulation.

When a female firefly produces a signal that helps males of her species locate her, this is plausibly a benign *other directed* epistemic action. Males find her more quickly or reliably with her flashing than without it. This exhibits a familiar pattern: many actions, including those in developmental and instructional settings, make sense as other-directed aids to cognition, and are standardly recognised as cognitive scaffolding for their targets. If we grant that the flashing of the conspecific firefly is an epistemic action, what of the flashing of the *Photuris*? Her signal works to the extent that it has similar effects on male fireflies of other species. It is just as much an other-directed epistemic action, and the same goes for beetles making food-eliciting gestures and cuckoo chick begging. There is a dramatic difference in the

consequences of being influenced because these epistemic actions by parasites are *hostile*, making cue-bound, agents *less reliable* in tracking the regularities their control systems evolved to track, and *more reliable* in benefiting the parasites.

We can't say that *all* cases of manipulation by cues are hostile epistemic actions, because in many cases the cues aren't produced by what the manipulator *does* but depend on non-behavioural properties of its phenotype. The cuckoo chick in Spain with its bright orange gape causes the birds it exploits to feed it more energetically, as does the cuckoo chick in England making more rapid calls. But one of them is more clearly behaving than the other, even if the one in Spain needs to display its compelling gape. There is no difficulty here for the strict extended geneticist, who isn't interested in the details of *how* the causal effects are constructed. We needn't do violence to the notion of an action and count all cases of parasite manipulation via cues as hostile epistemic actions but should rather recognise that not all mechanisms that produce deceptive cues are actions.[5]

Recognising some instances of manipulation by cue fabrication as hostile epistemic action brings the extended cognition and extended phenotype vocabularies closer together and illustrates the correctness of Sterelny's point that any model of cognition "that leaves out hostility is not a model of cognition in the wild" (2000:215). But it doesn't show how thinking about extended cognition can replace Dawkins' own cognitive internalism. I turn to that task now.

4.1 Exploiting distributed cognisers

Dawkins, recall, imagines that replicators must get at the *brain* or *nerves* of the organism they have an interest in influencing, whether through the front door or the service elevator. The phenomenon of distributed cognition, introduced in section (3) above, suggests a further possibility, which is manipulation *of extended cognitive processing*. If the functional stages of cognition can form chains that extend into the world, with significant cognitive work being done on the 'outside', then then those external stages might provide opportunities for manipulation. This would be distinct from the cases of manipulating by means of cues, which focuses on triggers that follow the normal sensory input channels in ways consistent with cognitive internalism. It would also be different from the processing-focused cases that Dawkins considers because it wouldn't require intervention to happen inside the nervous system. Some speculative examples will help flesh out the idea.

Imagine, first, that reed-warblers engaged in self-directed epistemic actions. Suppose that after every feeding trip in a day they place a marker in their nest and they stop collecting food for chicks when there are three markers. Then, first thing every morning, they remove the markers, thus resetting their external

[5] Behaviour and development are, of course, sometimes difficult to disentangle (see, e.g., Godfrey-Smith 2002).

memory. Such a scaffolded cuckoo would be more than "an active, complex machine, with sense organs, muscles and a brain" (Dawkins 1982:68). And its external cognitive resources would be vulnerable to manipulation by actions of a cuckoo chick. Such chicks are *already* disposed to eject reed-warbler eggs and hatchlings. Ejecting one marker whenever there were more than one would be a simple behavioural disposition that could drive additional feeding trips by the scaffolded host. Indeed, *any* host epistemic action or external cognitive resource within reach of the cuckoo chick and relevant to its parasitic mission would be a potential target for manipulation.

Now consider ant pheromone trails, which are real cases of scaffolding produced by non-human epistemic actions. The dispositions of ants to deposit markers, and to condition the direction of their own movement on the markers, mean that a route from a food source to the nest can be marked out and exploited by many individual agents that are not themselves particularly smart. An accumulation of pheromone deposits into a *trail* from nest to food source can 'carry information about direction and distance' and the intensity of a trail 'carries information about the value of the food resource' (Sterelny, 2003: 19). If we take this seriously the trail is a kind of structured representation, carrying information not possessed by the ants themselves, either individually or collectively. It is a temporary epistemic resource somewhat analogous to a navigation chart.[6] Once we recognise the trail as a distributed cognitive resource, we can speculate about exploitation and manipulation that targets it. A creature that fed ants to its young and could produce ant pheromones could lay a trail from an ant nest to where it kept its young. Or rivals for a food source could lay a stronger trail branching off an existing one, and which led nowhere long enough to exhaust the food.

Alternatively, consider the spider web not as an extension of the organs of a cognitively internalist agent, but as a cognitive tool. The strands of a web transmit vibrations. The web isn't passive because threads of different thicknesses respond differently and aren't randomly arranged. The web integrates and pre-processes information about the direction, distance, size, and vigour of things moving on it just as the architecture of sense organs does. The spider can modify the transmission properties of the web by selectively adjusting the tension of threads. The degree of attention given by a spider to a region of the web is sensitive to the tension, and the tensioning activity responsive to factors including hunger level, and experienced productivity of different web regions (Watanabe 2000; Nakata 2010). Such considerations motivate Japyassú and Laland (2017) to say that webs satisfy a mutual manipulability condition for 'constitutive relevance' as part of an extended cognitive apparatus.

Spiders that predate on web-builders exploit these factors. Some araneophagic jumping spiders manipulate the webs of other spiders in ways conditional on the size of the prospective prey (Jackson &

[6] I say this to highlight the trail's information carrying functions. A pheromone trail precisely isn't a chart to the extent that it isn't a separate representation, but rather a kind of overlay or tagging of the world itself.

Wilcox 1990 cited in Barrett 2011:71). They make vibrations that solicit a full attack from smaller spiders, and ones that invite inspection but not attack from larger ones (Tarsitano et al. 2000, cited in Japyassú and Laland 2017). Some attacking spiders accelerate their advances when wind shakes the web of the target spider, and any vibrations caused by the predator likely to ignored or not noticed against the general movement (Wilcox, Jackson & Gentile, 1996 cited in Barrett, 2011:59). It isn't hard to *imagine* more sophisticated interventions, such as a predator increasing or decreasing the tension in some regions of the target spider's web to bias the builder's attention and behaviour, or interfering with the web by adding or removing strands so that directional and other information is distorted. The more computational work we imagine the web doing, the more sophisticated we can imagine the manipulations to be. In these cases, the manipulation would be achieved by intervening in *external cognitive processes* of the target agent. The lesson here isn't to *deny* that the spider in its web isn't at the same time an iconic case of an extended phenotype and an analogy for the extended network of beneficial cognitive resources that an individual can assemble. Rather the point is to emphasise that extended cognitive structures and processes are also vulnerabilities and opportunities for the expression of hostility.

The speculations above are intended, perhaps unrealistically in some details, as cases of manipulation by extended phenotypes. It is easier to imagine exploitation by human actors because humans have so much cognitive technology, and because doing so allows planned or premeditated interventions. Suppose, for example, that your community made its living from raiding ships that ran aground in a particular place. One way to make more ships run aground in the right place would be, by some means, to make as many ships as you could have navigation maps which worked well enough for everything else, but which incorrectly showed un-obstructed lines of constant bearing near your rocky shallows. The operations of drawing a straight line on their maps would normally be reliable transformations of a complex navigation task as Hutchins (1995) explains but would occasionally and non-accidentally be misleading. Such maps would be tools for significantly distributed cognition that were benign for most navigation tasks, and hostile, perhaps lethally so, for a select few. We don't need to suppose that cartographically assisted piracy, or even navigation charts, are instances of extended phenotypes to see how extended cognition could be associated with exploitable vulnerabilities.

The extended phenotype perspective effectively makes a prediction. In the gene-centric case it is that if *any* mechanism of manipulation gains more for its genetic cause than it costs, and can be found by natural selection, *it probably will occur*. (And then perhaps precipitate an arms race.) Dawkins argues that we'll see more opportunities for, and understand more real cases of, manipulation *if we're extended functionalists about phenotypes*. In the narrowly biological cases the 'can be found by selection' condition comes with the usual constraint that every intermediate stage must pay its way. If we combine hostility with extended functionalism, but don't insist on gene-centrism, we should be open-minded both about the forms hostility can take, and where it might be found. Except for one of the speculations above, I've restricted myself to imagining forms of manipulation via distributed cognition that could be cases of

extended phenotypes. That helped bring out connections between thinking about extended phenotypes and extended cognition. Few if any defenders of extended cognition are, though, gene-centrists. In the following section I focus on extended hostility without gene-centrism.

5. Hostility without Gene-Centrism

If we adopt an interest-agnostic notion of, the prediction is that if *any* mechanism of manipulation gains more for its producer than it costs and can be found by whatever selection process is relevant, then *it will probably occur*. Here too, extended functionalism motivates us to be open-minded about where the means might be found as well as how they might work. The obvious candidates for producers besides genes include individual organisms and groups of individuals, including organised groups like corporations and states. (Some would include cultural replicators on this list.) A less obvious candidate, but worth emphasising, is the sub-systems or parts of agents, however extended they might be. As we saw above, the interests of genes don't always fully coincide with those of the organisms they occupy, and are sometimes expressed when the means can be found.[7] When cognitive sub-systems are themselves agent-like, responsive to local reinforcement, harmony between their goal-seeking and overall interests isn't guaranteed.

More generally, that is, the producers of manipulation could be anything with something to gain from it. The targets can in turn be any kind of agent worth exploiting. In human cases these might include relatives, friends, customers, competitors, employees, subjects of criminal justice or medical systems, the poor, the politically marginal, members of hated out-groups, and so forth. Next, the processes of selection needn't be restricted to vanilla natural selection but can include the possibly encultured capacities for learning and innovation of individual organisms, as well as organised research and development whether in the form of relatively disinterested 'science' or in corporate and other institutional environments. Extended phenotypes must be naturally selected, but hostility can be planned as well as refined through trial and error. All of this *makes things much worse* because some of those working to manipulate and exploit have substantial research and development capabilities. They can afford successions of prototypes that don't work (yet). They can unleash tools that use reinforcement learning at fearsome speeds to figure out how to pick our locks while we interact with them. They can run all manner of experiments piloting new versions of their products on subsets of their users, allowing their already profitable activity to cross-subsidise trials to find anything that works better. And, just like arms races between big slow organisms and small pathogens, they sometimes get to play rounds faster than our protections, mostly in our individual educational and collective political and regulatory processes, can keep up. (Education can arm against some attempts at manipulation, while regulation can transform the costs and benefits of being

[7] The suggestion that conflict between genes might lead to psychological conflict within an individual has been made by Haig (e.g. 2006) and Trivers (e.g. 2009). See Spurrett (2016).

caught doing it, or engaging in some kinds of research and development.) Finally, the manipulative *means* available to this larger set of actors with a wider range of selective processes can be as varied as extended functionalism allows. Sticking to those relevant to extended cognition, that includes cues, scaffolding, niches, cognitive tools, some cultural products, institutions and built spaces among things that might provide vulnerabilities. Any extended process could be a target, along with the familiar internal ones.

It is worth noting that Sterelny has said he *doesn't* think manipulation of public epistemic resources is likely. He grants that "hostile manipulation of [our] informational environment is a serious danger", but thinks the danger is restricted to "one-on-one high-stakes negotiations" (Sterelny, 2010:474), such as someone exploiting or harming the notebook-using Otto by erasing or altering his external memory. Sterelny argues that manipulation involving *public* resources (such as deceptively changing the maps in a subway station) is, in contrast, unlikely *because* sharing itself increases reliability, and the fact that many agents use different copies of the resources at unpredictable times makes it difficult to exploit a chosen target. Shared epistemic niches, that is, are protected against some manipulators.

Public epistemic resources are indeed often characterised by considerable redundancy, as in the case of underground train maps, in ways that mean it would be prohibitively expensive to manipulate such systems to misdirect a single passenger. The considerations that Sterelny emphasises aren't, however, fully general. In the imagined malicious navigation chart case at the end of section (4.2), for example, it might only be necessary to replace a single chart to influence a ship's course because ships at sea aren't like underground train customers in having many maps to choose from. In places like casinos, owners have comprehensive control over the entire environment, including its layout and design, and what activities are available where. The same goes for some retail spaces and, as Slaby (2016) notes some workplaces. Not only that, achieving similar levels of control over an environment to a casino boss doesn't require command over a large physical space because an increasing number of the environments people face are *virtual*, and control has only to be achieved within a single application. The potential victims of a variety of forms of exploitation carry powerful and almost permanently connected computing devices around with them in the form of their mobile phones. So, while the points Sterelny (2010) makes do show that some manipulation would likely cost more than it made and illustrate the protections that crowd-sourcing can bring, they don't defeat the general prediction that if *any* mechanism of manipulation gains more for its producer than it costs and can be found by whatever selection process is relevant, *it will probably occur*.

There are too many examples for anything but a selective and breathless survey here. Ross has pointed out that while elephants and baboons get drunk when they find low-toxicity sources of alcohol, such as fermented fruit, they "are at no risk of addiction […] because they cannot cultivate sources of low-toxicity alcohol. Their parties are windfalls, the frequency of which they cannot influence" (Ross, 2020:6). Human innovation has allowed stockpiling and the building of environments that facilitate addiction by allowing freer scheduling of consumption. Timms and Spurrett (MS) analyse some features

of electronic slot machines that present misleadingly frequent 'near miss' outcomes and in other ways support inaccurate estimation of the odds of winning as instances of *hostile scaffolding* that targets extended cognitive processes. Turning from customers to employees, Slaby (2016) has argued that we shouldn't assume that the affective scaffolding found in workplaces, and the habits it encourages and norms of interaction and behaviour it promotes, will reliably serve the interests of employees. Rather, he argues, those environments can effect a "hack" of the subjectivity of employees better characterised as "mind invasion" than mind extension. Alfano et al (2021), on the other hand, find evidence that the YouTube algorithm does seem to amplify extremist content and conspiracy theories. Google doesn't gain directly from any benefit to extremist groups, but it does gain from the increased advertising revenue when people watch extended series of videos. There are many more cases.

I've focused mostly on the importance of hostility for extended cognition here. But as noted a standard argument for extended *minds* begins by arguing for extended functionalism about cognition, and then goes on to argue that minds are where their realisers are. This is sometimes attended by complications involving specifying further criteria, such as reliable or fluent access (Clark & Chalmers 1998; Clark 2010) to distinguish parts of the mind 'proper' from relatively transient couplings with tools and resources that might extend cognition without extending minds. That being the case, the considerations here are *prima facie* relevant to extended minds, subject to decisions about those further issues.

Clark sometimes argues that there is a *moral* justification for taking extended minds seriously, which is precisely that if we do, we'll recognise the significance and severity of some of the harms extended cognisers can suffer (e.g. in the *Philosophy Bite*s discussion of 'The Extended Mind', also Clark 2008:232; Clark 2003:5). We only need to suppose that a fraction of our extended cognitive networks play a significant role in constituting our minds and selves to see something in this. And if we take hostility seriously, and the prediction that if there is something for anyone to gain, by whatever means, from hostile activity, or scaffolding, or technology, or niches, or institutions and so forth, we need to recognise that we're potentially up against an enormous variety of harmful mechanisms, especially as cheap and ubiquitous computing power, and the fact of enormous sharing of personal information in continuously networked environments, makes it possible for learning systems to fine tune increasingly bespoke forms of manipulation and exploitation. If some of the extended systems that are manipulated to exploit people are self-constituting, then the harms of being manipulated may indeed be more significant. This isn't predicted by the general principle that any discoverable means of exploitation that pays its way will probably occur because it gives no additional benefit to the exploiter unless malice is among their motivations. But cuckoos of many different kinds may be at work wherever our minds extend.

References

Aagaard, J. (2020) 4E cognition and the dogma of harmony. *Philosophical Psychology*, 34(2):165-181.

Ågren, J.A. (2021) *The Genes' Eye View of Evolution*. OUP.

Alfano, M., Fard, A.E., Carter, J.A., Clutton, P., & Klein, C. (2021) Technologically scaffolded atypical cognition: the case of YouTube's recommender system. *Synthese,* 199:835-858.

Alvarez, F. (2004) The conspicuous gape of the nestling common Cuckoo Cuculus canorus as a supernormal stimulus for Rufous Bush Chat Cercotrichas galactotes hosts. *Ardea*, 92:63–68.

Brooks, R.A. (1991) Intelligence without representation, *Artificial Intelligence*, 47:139–159.

Clark, A. (1997) *Being There: Putting Mind, Brain and Body Together Again*. Cambridge, Massachusetts: MIT Press.

Clark, A. (1998) Where Brain, Body and World Collide. *Daedalus*, 127(2):257-280.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58:7–19.

Clark, A. (2002) Towards a science of the bio-technological mind. *International Journal of Cognitive Technology*, 1(1):21–33.

Clark, A. (2003) *Natural Born Cyborgs*. Oxford University Press.

Clark, A. (2007) Curing Cognitive Hiccups: A Defence of the Extended Mind. *Journal of Philosophy*, 104(4):163-192.

Clark, A. (2008) *Supersizing the Mind*. Oxford University Press.

Clark, A. (2010) Memento's revenge: The extended mind extended. In R. Menary (Ed.), *The extended mind*. Cambridge: MIT.

Clark, A. (2012) Embodied, embedded, and extended cognition. In Frankish, K. And Ramsay, W.M. (eds) *The Cambridge Handbook of Cognitive Science*. Cambridge University Press:275-291.

Davies, N.B., Kilner, R.M. & Noble, D.G. (1998) Nestling cockoos, Cuculus canorus, exploit hosts with begging calls that mimic a brood. *Proceedings of the Royal Society B,* 265(1397):673–678.

Davies, N.B. (2015) *Cuckoo: Cheating by Nature*. Bloomsbury.

Dawkins, R. (1976b) Hierarchical organisation: a candidate principle for ethology. In P. P. G. Bateson & R. A. Hinde (eds), *Growing Points in Ethology*, Cambridge University Press:7-54.

Dawkins, R. (1982) *The Extended Phenotype*, Oxford: Oxford University Press.

Dawkins R. (2004) Extended phenotype–but not too extended. A Reply to Laland, Turner and Jablonka. *Biology and Philosophy*. 19:377–396.

Dawkins R. (2012). Foreword. In DP Hughes, J Brodeur, and F Thomas, (eds), *Host Manipulation by Parasites*, Oxford University Press:xi–xiii.

Godfrey-Smith, P. (2002) Environmental Complexity and the Evolution of Cognition. In Robert J. Sternberg and James C. Kaufmann (eds.) *The Evolution of Intelligence*, Lawrence Erlbaum:223-249.

Haig, D. (2006) Intrapersonal conflict. In: Jones MK, Fabian AC (eds) *Conflict*. Cambridge University Press, Cambridge:8–22.

Hutchins, E. (1995) *Cognition in the Wild*. Cambridge, Massachusetts: MIT Press.

Jackson, R.R., & Wilcox, R.S. (1990). Aggressive mimicry, prey-specific predatory behaviour and predator recognition in the predator-prey interactions of Portia fimbriata and Euryattus sp. jumping spiders from Queensland. *Behavioral Ecology and Sociobiology*, 26:111–119.

Japyassú, H.F., & Laland, K.N. (2017) Extended spider cognition. *Animal Cognition*, 20:375-395.

Kirsh, D. & Maglio, P. (1994) On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, 18:513-549.

Kirsh, D. (1995) The intelligent use of space. *Artificial Intelligence*, 73:31-68.

Krebs, J.R., & Dawkins, R. (1984[1978]) Animal Signals: Mind-Reading and Manipulation. In J.R. Krebs and N.B. Davies (eds.) *Behavioural Ecology: An Evolutionary Approach*, second edition. Blackwell.

Menary, R. (2010) The extended mind and cognitive integration. In R. Menary (Ed.), *The extended mind.* Cambridge: MIT.

Nakata, K (2010) Attention focusing in a sit-and-wait forager: a spider controls its prey-detection ability in different web sectors by adjusting thread tension. *Proceedings of the Royal Society B,* 277(1678):29–33.

Ross, D. (2020) Addiction is socially engineered exploitation of natural biological vulnerability. *Behavioural Brain Research, 386*.

Rupert, R. (2016) Embodied Functionalism and Inner Complexity: Simon's 21st-Century Mind. In R. Frantz and L. Marsh (eds.) *Minds, Models, and Milieux: Commemorating the Centennial of the Birth of Herbert Simon*. Basingstoke: Palgrave Macmillan:7–33.

Shapiro, L. (2004). *The mind incarnate*. Cambridge: MIT Press.

Spurrett, D. (2016) Does Intragenomic conflict predict Intrapersonal conflict? *Biology and Philosophy* 31(3): 313-333.

Sterelny, K. (2000). Roboroach: the extended phenotype meets cognitive science. *Philosophy and Phenomenological Research*, 61:207–215.

Sterelny, K. (2003) *Thought in a Hostile World*, Oxford: Blackwell.

Sterelny, K. (2010) Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences*, 9(4):465-481.

Sutton, J. (2010) Exograms, interdisciplinarity and the cognitive life of things. In R. Menary (Ed.), *The extended mind*. Cambridge: MIT.

Tarsitano, M., Jackson, R.R., & Kirchner, W.H. (2000) Signals and signal choices made by the araneophagic jumping spider Portia fimbriata while hunting the orb-weaving web spiders Zygiella x-notata and Zosis geniculatus. *Ethology,* 106(7):595–615.

Timms, R. & Spurrett, D. (MS) Hostile Scaffolding. https://philarchive.org/rec/TIMHS

Trivers, R. (2009) Genetic conflict within the individual. Berlin: *Sonderdruck der Berliner-Brandenburgische Akademie der Wissenschschaften*. 14:149–199.

Watanabe T (2000) Web tuning of an orb-web spider, Octonoba sybotides, regulates prey-catching behaviour. *Proceedings of the Royal Society B,* 267(1443):565–569.

Wheeler, M. (2010) In defence of extended functionalism. In R. Menary (ed) *The Extended Mind*, MIT Press:245-270.

Wilcox, R.S., Jackson, R.R., & Gentile, K. (1996). Spiderweb smokescreen: spider trickster uses background noise to mask stalking movements. *Animal Behaviour*, 51:313–326.

Acknowledgments: An earlier version of this material was presented at the Philosophy of Biology at Dolphin Beach conference in July 2022. I'm grateful to Ron Planer, Rachel Brown, Kim Sterelny, and Peter Godfrey-Smith for their questions and comments. I also presented this to the department of Philosophy at Stellenbosch University in November 2022, and thank JP Smit for his engagement. Anita Craig read an earlier draft and helped me improve the exposition. I'm also grateful to John Sutton for comments and feedback.