Authors:
Zoe Drayson and Andy Clark

Title:
Augmentation, agency, and the spreading of the mental state

Affiliations:
Zoe Drayson, Department of Philosophy, University of Bristol, 9 Woodland Road, Bristol BS8 1SH, UK
Andy Clark, Department of Philosophy, School of Philosophy, Psychology and Language Sciences (PPLS), Dugald Stewart Building, Edinburgh EH8 9AD, UK

Corresponding author (Zoe Drayson):
zoe.drayson@bristol.ac.uk
Phone: +44 (0)117 928-7825
Fax: +44 (0)117 928-8626

Abstract:
According to the thesis of the ‘extended mind’ (Clark and Chalmers (1998)), the physical systems that realize some of our mental states and processes are not always confined to the brain. Neil Levy (2007a, 2007b) supports the extended mind thesis, and argues that it has ethical implications which have thus far been largely overlooked. We agree with Levy, but we emphasise the difference between the claim that the mind is extended and the claim that the mind is merely embedded, and suggest that the former may be associated with more significant ethics implications than the latter. We consider how the extended mind thesis has implications for strategies of cognitive rehabilitation, and for our understanding of cognitive impairment itself. We conclude with an exploration of the relationship between the boundaries of the mind and the boundaries of moral agency.

Keywords:
Cognition, ethics, extended mind, rehabilitation
Augmentation, agency, and the spreading of the mental state

Zoe Drayson and Andy Clark

1. Introduction

Scientists and clinicians can interfere with our brains as never before: with a heady mix of pharmaceuticals and technology, they can manipulate our moods and polish our powers of concentration. Do these advances in neuroscience bring with them a host of new ethical problems? It doesn’t seem obvious that new technology should automatically mandate the creation of a dedicated ethical field: despite the advances in heart surgery over the past decade, we don’t have a special discipline of ‘cardioethics’. But there is a widely held intuition that there is something special about the brain which gives rise to particular ethical considerations, and this intuition is responsible for creating the discipline of ‘neuroethics’. Researchers in neuroethics develop ethical frameworks for regulating neuroscience and its application to human beings [1].

What’s so special about the brain? The idea is, of course, that the brain is uniquely associated with our mental capacities, and that it is our mental capacities (not our heart or our lungs) that make us who and what we are: it is our mental capacities that both sculpt and realize our beliefs, our desires, our moods, and our skills. If advances in neuroscience thus allow for the manipulation of the very things that make us the human beings we are, then this does indeed seem worrying, and able to mandate the creation of a special sub-field of ethical theory. But are these worries anything new? Neil Levy [2, 3] suggests not. He points out that new brain technology only raises new ethical questions if it provides a unique way to manipulate mental states. At this point it is important to note that Levy supports what has become known as the thesis of the ‘extended mind’ [4], according to which the physical systems that realize some of our mental states and processes are not always confined to the brain. Instead, they may at times extend beyond skin and skull, recruiting notebooks, electronic resources, and other forms of bio-external ‘scaffolding’
as quite literally parts of the machinery of a human mind. Seen in this light, advances in neuroscience may give us new methods of intervention and manipulation, but it is not clear that they raise brand new ethical issues: if minds are already more than just brains, then neural intervention looks far less special, and would represent merely one way among many in which we may 'directly' interfere with mental states.

In section two, we set out this idea of the ‘extended mind’. We then (section 3) introduce Levy’s account of the implications of the extended mind thesis for the discipline of neuroethics, taking care to re-emphasize an important distinction between the idea of the mind as genuinely extended and the idea of the mind as merely embedded in a web of enabling structure. In section four we explore the concept of cognitive rehabilitation, and consider how the extended mind thesis has implications for rehabilitative strategies. This leads us (section 5) to a discussion of the nature of cognitive impairment itself, where we explore what enables some people to function capably despite brain damage. Despite some important points of disagreement, we conclude that the extended mind thesis does indeed offer an enlightening way to look at current developments in neuropsychology, and is at least suggestive of a new take on the ethical issues surrounding both cognitive function and dysfunction. We conclude with an exploration of the relationship between the boundaries of the mind and the boundaries of moral agency.

2. Extending the Mind

Proponents of the so-called 'extended mind' [4, 5] hold that even quite familiar human mental states (such as states of believing that so-and-so) can be realized, in part, by structures and processes located outside the human head. Such claims go far beyond the important but far less challenging assertion of potent cognitive embedding: the claim [6] that human cognizing leans heavily on various forms of external scaffolding and support. Instead, extended mind theorists paint mind itself (or better, the physical machinery that realizes some of our cognitive processes and mental states) as, under humanly attainable

---

1 Some of the material in this section is drawn from Clark (2003) and Clark (2008), by permission.
conditions, extending beyond the bounds of skin and skull. The machinery of mind, if this is correct, is not simply the bio-machinery contained within the ancient skinbag. In the present section we briefly rehearse the main argument before proceeding to examine its consequences for the special domain of neuroethics.

Clark and Chalmers’ [4] aim was to show that external traces (such as pencil marks in a notebook) may sensibly be considered, given the right additional circumstances, as among the physical vehicles of specific dispositional beliefs. This occurs if the traces become poised for the control of action in roughly (the 'roughly' is important, as we shall later see) the same kind of way as internal memory traces, yielding an extended material base for some of the agent's dispositional (i.e. genuine, but not consciously occurant) beliefs.

The claim here was not, implausibly, that an external, passive, encoding might somehow behave exactly like the fluid, automatically responsive resources of internal biological memory. Rather, it was that external encodings were, under certain circumstances, capable of becoming so deeply integrated into online strategies of reasoning and recall as to be only artificially distinguished from proper parts of the cognitive engine itself.

At the heart of arguments for the extended mind lies what has become known as the Parity Principle:

\[
\text{PP} \\
\text{If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process.} \\
\text{(from [4], p.8)}
\]

In other words, for the purposes of identifying the material vehicles of cognitive states and processes, we should ignore the old metabolic boundaries of skin and skull, and
attend instead to the computational and functional organization of the problem-solving whole. The parity principle thus provides a 'veil of ignorance' style test meant to help avoid bio-chauvinistic prejudice concerning the nature and location of the machinery of mind.

To show the principle in action, Clark and Chalmers offer a thought experiment. This is the case of Otto and Inga. Inga hears of an intriguing exhibition at MOMA (the Museum of Modern Art in New York). She thinks, recalls it's on 53rd St, and sets off. Otto suffers from a mild form of Alzheimer's, and as a result he always carries a thick notebook. When Otto learns useful new information, he always writes it in the notebook. He hears of the exhibition at MOMA, retrieves the address from his trusty notebook and sets off. Just like Inga, we claimed, Otto walks to 53rd St. because he wants to go to the museum and believes (even before consulting his notebook) that it is on 53rd St. The coarse functional poise of the stored information is, in each case, sufficiently similar (Clark and Chalmers argue) to warrant similarity of treatment. Otto’s long-term beliefs just aren’t all in his head.

A natural objection is to suggest that:

“all Otto actually believes (in advance) is that the address is in the notebook.
That’s the belief (step 1) that leads to the looking (step 2) that then leads to the (new) belief about the actual street address”

Call this the Otto 2-step. Despite its initial plausibility, the objection fails. Suppose we now ask why we do not depict Inga in similar terms? Why don’t we say that Inga's only antecedent belief was that the information was stored in her memory, and depict her retrieval as an Inga 2-step? Intuitively, the reason seems to be that in the case of Inga, the 2-step model adds spurious complexity: "Inga wanted to go to MOMA. She believed that her memory held the address. Her memory yielded 53rd St.”. What’s more, it seems likely that in the normal course of events Inga relies on no beliefs about her memory as such. She just uses it, transparently as it were. But ditto (we may suppose) for Otto: Otto is so used to using the book that he accesses it automatically when bio-memory fails. Calls to the notebook are deeply and sub-personally integrated into his problem-solving
routines. The notebook has become 'transparent equipment' for Otto, just as biological memory is for Inga. And in each case, doesn't it add needless and psychologically unreal complexity to introduce additional beliefs about the book or biological memory into the explanatory equations?

Overall, then, the claim was that Inga’s biological memory systems, working together, govern her behaviors in the ways distinctive of believing and that Otto’s smeared-out bio-technological matrix (the organism and the notebook) governs his behavior in many of the same kinds of way. So the explanatory apparatus of mental state ascription gets a grip in each case and what looks at first like Otto’s action (looking up the notebook) emerges as part of Otto’s thought.

The extended mind model thus claims that there is no watershed difference between information that is stored in the biological brain and information that is stored in non-biological structures. It is properties of organization and accessibility that matter, not the further details of exactly where and how information is stored.

3. Levy on the extended mind

Neil Levy [2, 3] supports the extended mind thesis, and argues (correctly in our view) that it has ethical implications which have thus far been largely overlooked. In particular, Levy rejects the idea that advances in neuroscience present brand-new ethical problems. He suggests that if the extended mind thesis is true, then the manipulation of minds does not require intervention into brains; if our mental states and processes can be partly realized by aspects of the environment, then they’re already publicly accessible. The new technology and pharmacology that allows interventions into people’s brains doesn’t necessarily offer anything distinctive: if we think that being able directly to access people’s minds raises ethical issues, then these ethical issues have been around for a long time.
Levy develops his argument using a version of the parity principle from Clark and Chalmers:

“our new ways of altering the mind are not, for all that, entirely unprecedented, and ought not to be regarded, as a class, as qualitatively different in kind from the old… This is not to say that they do not present us with genuine ethical dilemmas and with serious challenges; they do. But, for the most part, these dilemmas and challenges are new versions of old problems.” (from [2], p.xii)

Above, we claimed that proponents of the extended mind go beyond the claims of mere ‘embeddedness’: they do not just claim that our mental abilities are to some extent dependent upon environmental support, but that the mental states and processes themselves extend beyond the brain and body. The mind is merely embedded rather than extended, suggests Robert Rupert, when it depends heavily on the structure of the organism’s external environment but the external props are not seen as constituting part of the processing [6]. If minds are embedded, then we can only understand them by taking into account how they manipulate and exploit the environment. But if minds are extended, then the organism and its environment create an interactive system that is ‘most fruitfully viewed as a single unified system’ ([6] p.394). Rupert argues that even if it’s true that we cannot fully understand minds unless we consider the context in which they are embedded, this does not imply that the embedding context must be part of the mind itself.

Levy acknowledges the difference between the embedded mind and the extended mind when he distinguishes between a weak and strong form of what he labels the ‘ethical parity principle’ ([2], p.61). Like the original version of the parity principle found in Clark and Chalmers [4], Levy’s version is concerned with when it is appropriate to consider internal (neural) and external processes as equivalent in some sense. Levy’s ethical parity principle (EPP) considers when alterations of external cognitive props are ethically equivalent to alterations of the brain. The strong version of the EPP, which Levy supports by reference to the extended mind thesis, claims that all alterations of external cognitive props are ethically on a par with alterations of the brain. The weaker version of
the EPP, which Levy supports by reference to the embedded mind thesis, claims that alterations of external props are ethically on a par with alterations of the brain to the extent to which our reasons for finding alterations of the brain problematic are transferable to alterations of the environment in which it is embedded.

Levy states ([2] p.61) that although he endorses the extended mind thesis, his arguments rely solely on the weaker version of the EPP. Thus his examples are mostly cases of merely embedded minds, rather than extended minds. Much of Levy’s argument is based on the claim that causal routes that are largely external to the mind are not, in any ethically relevant way, different from causal routes which are largely internal. But this can indeed be held without having to subscribe to the extended mind thesis. Levy writes that:

“Whether the mind is merely embedded, or actually extends into the environment, doesn’t much matter in the end – not once we see the extent to which thought, wherever it is located, depends upon the world.” (from [2], p.60)

But we are not yet convinced of this. It seems to us that that there remains a clear difference between mental processes depending on the world in the sense of requiring inputs from the world, or being about the world, and mental processes ‘depending’ on the world in the sense of the processes themselves being partially physically realized by non-neural matter. We agree with Levy’s claim that the extended mind thesis has ethical implications which have largely gone unconsidered, but we think that these ethical implications may be different – and generally more significant – than those associated with the merely embedded mind thesis.

4. Cognitive rehabilitation

Consider the case of cognitive rehabilitation, the process whereby someone’s impaired capacity to process and use information is improved, so as to allow increased functioning
in everyday life [7]. Cognitive rehabilitation falls into two main categories, restorative and compensatory [7]:

Restorative rehabilitation aims to restore the function of the damaged neural area or the impaired network. It relies on the neural plasticity, the natural abilities of the brain to ‘rewire’ itself. Restorative techniques for memory loss, for example, might include training the patient to remember longer and longer lists of facts in an attempt to build or strengthen neural connections.

Compensatory rehabilitation aims to achieve the same functional results as restorative rehabilitation, but in a different way. Rather than trying to ‘mend’ the brain damage, compensatory rehabilitative techniques use a variety of cognitive strategies (internal and external) to help the patient adapt to the presence of the brain impairment. Compensatory techniques for memory loss, for example, might include training the patient to use internal mnemonics or associative imagery, or external aids such as notebooks or pagers, or environmental structuring on a larger scale.²

Some have questioned which of these strategies should be the main focus of rehabilitation: restoring a damaged cognitive function, or developing compensatory or alternative ways of performing a task [8]. Some neuroscientists seem to think that as we learn more and more about the brain, all rehabilitation will be restorative, and compensatory strategies will be unnecessary. Coltheart [9], for example, appears to believe that models of neural functioning from cognitive neuropsychology are sufficient to plan cognitive rehabilitation. The idea behind such thinking appears to be that compensatory techniques are a useful substitute in cases where we lack either the neuroscientific knowledge or the neurosurgical technology to restore the damaged brain networks. External compensatory aids (for example the use of photos and labels),

² Note that these two categories can overlap: it is possible for compensatory strategies to have a restorative effect; both methods require learning and depend upon repetitive activation of associated cognitive processes [7].
according to this way of thinking, may provide helpful additional inputs to the damaged cognitive processes.

According to the extended mind thesis, however, external compensatory aids aren’t necessarily merely a fall-back method of rehabilitation, useful when the damaged areas of the brain can’t be restored. If they are poised for the control of action in the right way (see section two) then they can come to count as part of the agent’s own cognitive and control processes, despite being non-neural.

One incident that helps illustrate this possibility was revealed by a chance encounter between one of us (Clark) and Carolyn Baum, then the head of occupational therapy at the Washington University School of Medicine in St Louis, Missouri. When Clark explained his ongoing work on the extended mind Baum immediately warmed to the theme. Exactly this lesson, she felt, was emerging from her own work with a sub-population of inner-city Alzheimer’s sufferers in St. Louis. These patients were a puzzle because although they still lived alone, successfully, in the city, they really ought not have been able to do so. On standard tests (such as the CERAD protocol) they performed rather dismally. They should be unable to cope with the demands of daily life. What was going on?

A sequence of visits to their home environments provided the answer. These home environments, it transpired, were wonderfully calibrated to scaffold these biological brains. The homes were stuffed full of cognitive props, tools and aids. Examples included: message centers where they stored notes about what to do and when; photos of family and friends complete with indications of names and relationships; labels and pictures on doors; ‘memory books’ to record new events, meetings and plans; and "open-storage" strategies in which crucial items (pots, pans, chequebooks) are always kept in plain view, not locked away in drawers.

This real-life example highlights the way the extended mind thesis gives us a way of looking at cognitive rehabilitation whereby external aids take on a new importance. In
fact, the success of compensatory rehabilitative techniques involving external aids supports the idea that factors outside the brain feature heavily in cognitive functioning. External rehabilitation aids have become increasingly effective in recent years with the development of new technology, and individuals can now carry around devices which are discreet and portable, and which are interactive and adaptable. Compensatory approaches using environmental support are extremely successful at improving function: for attention-impaired patients, for example, using a portable pager to cue important activity can induce performance which is indistinguishable from controls [10]. This contrasts with rehabilitative approaches which regenerate the central nervous system: they may successfully restore the damaged neural machinery, but the extent to which regeneration leads to functional gains in coping with real-life problems is less clear [11].

The extended mind model thus invites us to view cases of potent non-biological 'scaffolding' as implementing alternative but perfectly genuine forms of control and recall, rather than viewing them merely as compensatory props and patches enabling better performance. Distinctions between restoration and compensation need to be based not on superficial facts about what is or is not a case of actual biological repair, so much as upon the details of the capacities that the new structures (be they biological or otherwise) hold in place.

Taking the extended mind story seriously requires us to reconsider many of our views and prejudices concerning cognitive rehabilitation and the understanding and depiction of cognitive impairment. For example, the forceful re-location of a functioning Alzheimer’s patient into a controlled hospital setting often constitutes a tragic turning point. Seen from the perspective of the extended mind, such re-location could be akin to the infliction of new brain damage upon an already compromised host. But as a society, we do not yet enjoy a structure of laws and social policies that recognizes the deep intimacy of agents and their cognitive scaffoldings: intimacy such that certain harms to the environment can simultaneously be harms to the person.
5. Cognitive impairment and ‘cognitive reserve’

The previous section highlighted one way that the extended mind thesis may have implications for the way we understand cognitive rehabilitation and in particular the relationship between the cognitively impaired person and their rehabilitative aids. In this section, we will argue that it is not just rehabilitation that looks different given this view, but also cognitive impairment itself.

Consider the following case study, based on Archer et al. [12]. A 73-year-old man – call him John – visited the National Hospital for Neurology and Neurosurgery in London. He was concerned that his ability to play chess had declined over the past two years: he would previously plan his game seven moves in advance, but now he was only able to plan three or four moves in advance. His family were not aware of him showing any sign of cognitive impairment, and there were no changes in personality or language, and he continued to look after the household finances. He scored normally on all the neuropsychological tests. John’s condition hadn’t changed at all by the time of his death three years later. But when the autopsy was carried out, John’s brain was discovered to have all the signs associated with advanced Alzheimer’s disease: severe neurofibrillary tangle pathology, cerebral amyloid angiopathy, and Lewy body pathology in the brainstem and limbic structures. This was unexpected, because most people with such severe brain degeneration would have lost many of their mental abilities and wouldn’t be able to function without lots of help and support. But John’s only problem seemed to be that his chess game wasn’t as good as it used to be.

John’s case is surprising, but not unique. For twenty years or so, we’ve known that people with the same degree of apparent brain damage can show very different symptoms [13-15]. Some people seem to have a sort of ‘mental padding’ or a ‘buffering’ capacity, which protects them from showing the most serious effects of brain damage. Neuropsychologists [16-18] have introduced the concept of ‘cognitive reserve’ to refer to this capacity for resilience that people like John exhibit. The higher a person’s level of cognitive reserve, the less likely they are to show the mental decline associated with age,
head injury, stroke, HIV, Parkinson’s disease, Alzheimer’s disease, or poisoning with neurotoxins.

But what is it that protects some people from the most devastating effects of brain degeneration? In other words, what is the mechanism underlying the phenomenon of cognitive reserve? Research shows that the key mediating factors associated with cognitive reserve are high levels of educational and occupational attainment, high IQ and literacy, and a stimulating and social lifestyle. The standard theory behind cognitive reserve is that these factors encourage new nerve cells to form in the brain (neurogenesis) and boost the brain’s capacity to reorganize itself as a result of experience (neural plasticity), so that even when parts of the brain are damaged, it can still function reasonably well. This interpretation of cognitive reserve postulates that “specific mental stimuli and challenges during both childhood and adulthood stimulate the formation of complex neuronal networks and promote cognitive reserve capacity” [19]; and that “neuronal plasticity permits cognitive reserve to be enhanced or maintained during the adult years” [20]. A recent review of neuroimaging studies of cognitive reserve concluded that the results “represent first steps in identifying the neural implementation of the concepts of reserve” [21]. According to the standard theory of cognitive reserve, if we want to explain how it is that people like John can lead normal lives despite the damage to their brains, then we should study the brain more carefully. The idea is that whatever is responsible for this phenomenon will be found inside our heads.

But is it really so clear that the explanation of cognitive reserve must rely solely on the brain? Think about the similarities between John and one of the Alzheimer’s patients in St Louis: they both have damaged brains and we wouldn’t expect them to function well in everyday life. But they do function well, leading reasonably independent lives. We can explain the abilities of the St Louis patient by the structure of her immediate environment, her support networks, and the way she uses gadgets and tools to help her. We don’t ask what it is about her brain that allows her to function so well despite the damage, so why do we think about John’s case this way?
One obvious difference is that we can immediately see how much help the Alzheimer’s patients in St Louis are getting from their environmental props. John, on the other hand, didn’t have any of these supports: his family reported that he was functioning as normal. Does this entitle us to conclude that the only explanation of John’s abilities is internal, and directly related to the functioning of his brain? We think not: external props aren’t always as obvious as those found in the St Louis house. Think of how many of our mental abilities seem to rely on factors external to the brain: counting on our fingers, for example, or taking a shopping list to the supermarket. The abilities of our own normal brains are limited – most of us find complex arithmetic difficult without a pen and paper, and we often write notes for ourselves to make up for our limited memories. Simple tools and props like these give us more complex abilities than our brains alone could. Even undamaged brains can benefit from external help sometimes.

Recall next that people with high cognitive reserve have often led busy lives with complex jobs and stimulating hobbies. Such people have probably developed a number of ways to cope with high everyday demands on them, which might include recruiting, as a matter of course, many external props and aids into their ‘normal’ routines. It’s possible that people who rely heavily on such tools and props in everyday life could find it easier to adapt to some forms of brain damage, because they’re already supplementing their brain’s abilities with external aids. The damage might only become functionally salient in artificial conditions, such as those enforced by the rules and practices of chess.

6. Concluding Thoughts: The boundaries of the agent

All this serves to highlight the complicated relationship between minds, persons, individuals, and agents. If the mind extends, does it follow that the person or the agent or the individual who has the mind extends likewise? The answers to this grammatically ugly question are interesting in their own right, but particularly pressing when discussing ethical issues. The question of morality arises with respect to mental interventions because having a mind generally denotes having other properties: consciousness,
rationality, personhood, or agency, for example. But it is not clear that the extended mind thesis has direct implications for any or all of these features: if we identify the agent as the combination of body and biological brain, then the boundaries of agent still stop at the skin. Clark [22] suggests that the “extension of the morally resonant notions of self, mind, and agenthood to include aspects of the world beyond the skin” is not a necessary implication of the extended view. Clark and Chalmers [4] argue only for the extension of cognitive processes and states, and are not committed to the idea that the material foundations of conscious experience extend beyond brains (for the opposing view, see Noë (2009) [23]. A similarly conservative view is held by Wilson [24], who supports the extended mind thesis, but denies that this has implications for the boundaries of the individual or the agent. Wilson suggests that the agent is the individual who bears the mental states, where that individual is a spatio-temporally bounded, cohesive and continuous organism, and the locus of causation or action.

If the interventions and manipulations discussed in neuroethics are seen as subjects for ethics debate because they are interventions into and manipulations of (the material bases of) persons or agents, then whatever implications one takes the extended mind thesis to have for neuroethics may be further mediated by one’s views of personhood and agency in their various guises. The extended mind thesis thus raises new ways of looking at the relationship between embodiment, individuality, and agency, while still leaving many important (and potentially morally salient) issues unresolved.

Acknowledgments

This paper was prepared thanks to support from the Arts and Humanities Research Council, under the European Science Foundation Eurocores CNCC scheme, for the CONTACT (Consciousness in Interaction) project, AH/E511139/1.

References


