Published in *Philosophy, Psychiatry & Psychology* (2011) 18: 339–344.
doi:10.1353/ppp.2011.0057

mark.sprevak@ed.ac.uk

Neural sufficiency, reductionism, and cognitive neuropsychiatry

Mark Sprevak
University of Edinburgh

11 January 2012

Kanaan and McGuire (2011) elegantly describe three challenges facing the use of fMRI to uncover cognitive mechanisms. They shows how these challenges ramify in the case of identifying the mechanisms responsible for psychiatric disorders. I want to raise another difficulty for fMRI that also appears to ramify in similar cases. This is that there are good reasons for doubting one of the assumptions on which many fMRI studies are based: that neural mechanisms are always and everywhere *sufficient* for cognition. I suggest that in the case of the mechanisms underlying psychiatric disorders, this assumption should be doubted. I do not dispute that a malfunctioning neural mechanism is likely to be a necessary component of a psychiatric disorder—as Kanaan and McGuire say, the experimental evidence from cognitive neuropsychiatry gives us excellent reasons to think that this is so. My question is whether a story *only* in terms of these neural mechanisms is sufficient to explain the mechanism of a psychiatric disorder. Is the reduction, projected by cognitive neuropsychiatry, of psychiatric disorders to disorders in neural functioning even in principle possible? Drawing on recent concerns about the location of mental states, I argue that such a reduction is likely to fail. Even if the considerable problems raised by Kanaan and McGuire for fMRI could be addressed, we have no reason to think that the mechanisms involved in psychiatric disorders are entirely neural, and that fMRI, or even a perfect science-fiction brain-scanner, would be capable of uncovering them. Psychiatric disorders, like numerous other cognitive processes, are liable to cross the brain-world boundary in such a promiscuous way as to be resistant to neural reduction.

As Kanaan and McGuire argue, part of the appeal of fMRI to psychiatry is that it offers the tantalising prospect of putting psychiatry on a firm biomedical foundation: *viz.* mechanistically explaining psychiatric disorders in terms of functional brain disorders. In the ideal case, this would involve finding a neural mechanism responsible for, sufficient for, or characteristic of, a given psychiatric disorder. A psychiatric patient could, say, be diagnosed with one or more malfunctioning cognitive mechanisms, which ultimately would be identified with a neural malfunction. The proposed identification would run in two steps. Mental disorders would first be characterised within the framework of cognitive psychology: in terms of malfunctioning cognitive, perceptual, behavioural, systems. Then those systems would be identified with underlying brain structures and functions responsible for their action via imaging studies such as fMRI. Thus, fMRI would link our existing cognitive and behavioural characterisation of mental disorders (patients that act, feel, or think in dysfunctional ways), with underlying neural mechanisms. If such a mechanistic reduction could be achieved, it would have the potential to dramatically increase our understanding of the nature of mental disorders. Psychiatric disorders would be understood as medical disorders, continuous with other bodily disorders studied in the biomedical sciences.

Kanaan and McGuire raise three challenges for fMRI as a way of uncovering neural mechanisms underpinning psychiatric disorders. These concern problems with task choice, statistical power, and interpretation of fMRI results. Task choice is the problem that, due to the nature of psychiatric disorders, it is hard to get a meaningful or sharp contrast between 'ill' and 'normal' behaviour from patients in an fMRI scanner. Statistical power is the problem that psychiatric disorders tend to involve fMRI measurements with a low SNR, and small or heterogeneous groups of patients, which produces results of dubious statistical significance. The interpretation problem is that scans of psychiatric patients are open to a variety of interpretations, but unlike fMRI scans of normal subjects, the assumptions relied on to select the correct interpretation may not hold true in psychiatric cases. Given these three problems, fMRI is far from a straightforward tool for identifying the neural mechanisms involved in a psychiatric disorder.

At least as we would wish it, the logic of fMRI is that it allows one to see the regions of the brain *sufficient* for a given cognitive function. The lesson from Kanaan and McGuire is that in practice fMRI can rarely deliver a clear or unequivocal answer to the sufficiency question. I wish to press a different source of worry: whether it is even possible for a (perfect) brain-scanner to reveal the mechanisms sufficient for a psychiatric disorder. The inference from an fMRI scan to a mechanism for a given cognitive function depends on the *neural sufficiency* assumption. This is the assumption that neural activity (of some kind or other) is sufficient for the cognitive process or state in question to occur. The neural sufficiency assumption guarantees that the mechanism involved in a cognitive function lies within the neural domain. If the mechanism underpinning that cognitive function were to fall in part outside that neural domain, or if it were to involve interactions between the neural and non-neural environment, then fMRI (or any other brain-scanner) would be unable to capture it.

In recent years, there has been increasing doubt that neural sufficiency is always and everywhere true for mental states and processes. The mechanisms underlying at least some mental processes appear ineliminably to involve the way in which the brain couples to certain props and features in the environment. These cognitive mechanisms do not appear to be explicable in internal neural terms. I wish to suggest that at least some of the cognitive mechanisms involved in mental disorders fall under this type. Hence, even if Kanaan and McGuire's worries about fMRI could be addressed, an in principle barrier to using a brain-scanner to identify the mechanisms of psychiatric disorders may exist, and hence a barrier to effecting the reduction of psychiatric disorders to neural disorders envisioned by cognitive neuropsychiatry.

Hurley (1998) presents a tempting picture of the mind she thinks we should resist called the *Input-Output Picture*. The Input-Output Picture understands cognitive activity as roughly a linear flow in which the senses deliver input, which is progressively processed by perceptual and cognitive resources, resulting in an output (usually a motor action), and then the whole process repeats. This picture views cognitive mechanisms as being essentially located *after* sensory input and *before* motor output. Because sensation and motor activity lie at neural interfaces, it seems natural to assume that cognitive mechanisms must lie somewhere in the neural circuitry between: sensation and motor activity bookend the neural boundary, and cognition is an intermediate step, so presumably cognition must be neural. What else *could* cognition be other than a neural process? And if cognitive mechanisms are neural mechanisms, then to the extent psychiatric disorders are cognitive disorders, they are naturally understood as disorders in neural functioning.

A number of objections have appeared to the Input-Output Picture and the neural sufficiency assumption that it motivates.[1] A common thread in these objections is that to construe the environment, as the Input-Output Picture does, as mere input to, or output of, a cognitive process seriously underestimates the role of environmental

[1] Concerns arise from a variety of perspectives and cognitive domains, see Clark (2008); Clark and Chalmers (1998); Dennett (1996); Haugeland (1998); Hurley (1998); Hutchins (1995); Menary (2007); Noë (2004); Rowlands (2003, 2006); Sprevak (2009, 2010); Wheeler (2005); Wilson (2004).

props in cognitive mechanisms. Dynamic feedback and feedforward loops run through the agent's perceptual, cognitive, motor systems, body and worldly props, and a description of how the entire loop behaves over time is often necessary to identify a cognitive mechanism. Cognitive agents actively structure their world, and those structures play a crucial role in their cognitive processes, which in turn guides further structuring. In many cases, the environment is not just an input; it is an essential part of the cognitive processing. Neural mechanisms combine, often in unexpected ways, with environmental props to get the job of cognition done. The mechanism involved often cannot be adequately described purely in terms of brain-side activity plus a specification of input and output. It may be difficult or impossible to explain the cognitive mechanism unless one tells a story that shows how neural resources and the environment couple together to achieve successful cognition. A brain-side story alone simply doesn't capture the cognitive mechanism.

Here is an example of the intended explanatory contrast. Consider the task of catching a fly ball in baseball (Clark 2008). This task involves what appears to be a cognitive/perceptual achievement: *viz.* working out where to stand to catch the ball. One might seek to explain this cognitive achievement in terms of internal mechanisms that make the relevant calculation of where to stand: e.g. neural mechanisms that predict the ball's future trajectory based on its observed position and velocity. In this case, the cognitive competence would be explained purely in terms of internal neural mechanisms. However, it turns out that this form of explanation, at least in this instance, is wrong: the mechanism involved in successfully catching a baseball is not wholly neural. The mechanism involves neural activity working together with the environment in a loop. The mechanism appears to be as follows: a fielder runs so that her optical image of the ball presents a linear constant speed movement against her visual field (McBeath, Shaffer and Kaiser 1995). This exploits an invariant in the optic flow, and by following this rule, the fielder is sure to arrive at the right place to catch the ball. Her mechanism for working out where to stand to catch the ball involves a combination of neural resources, bodily motion, and environment working together over time. While the first explanatory model assumed that what lay behind our cognitive achievement was an internal mechanism, and in principle we could find that mechanism somewhere in the brain (if only we could fit a brain-scanner to a moving outfielder!), on the second model, the cognising involved has to be explained in terms of a wider loop involving brain, body, and world. A brain-side story alone is insufficient to explain the mechanism underpinning solution of the cognitive task.

Another example is how players select their words in Scrabble (Clark and Chalmers 1998; Kirsh 1995). Players use physical re-arrangement of their letter tiles to prompt word recall during play, which in turn prompts further re-arrangements of the tiles, prompting further word recall, etc. This mechanism achieves a cognitive competence that is not available by purely internal thought. Here, the player's deliberation in choosing words is not just a matter of internal cognition. The deliberation process spills out into the world to include the player's interaction with the physical letter tiles. The tiles are not just net inputs or outputs to an internal deliberation process, they are part of the processing mechanism. This mechanism could be divided into neural and non-neural parts, but that may not be particularly desirable or helpful in explaining how the cognitive task gets done. Appeal to the whole mechanism is typically the best way to explain the player's cognitive competence. A brain-scanner, even if attached to a player during play, would not be able to reveal the mechanism behind her word choice.

A natural thought one might have about these cases is to insist that only the neural part of the cognitive processes deserves to be called the 'cognitive mechanism', and everything else should be understood as mere input or output.[2] This move deserves more consideration than can be given here, but one immediate problem it faces in this context is that it cuts against the explanatory interests that motivate talk of cognitive mechanisms in the first place. Typically, we want an explanation for why certain behaviour, thoughts, beliefs tend to occur: a mechanism for how they are generated. This is, perhaps above all else, the motivation for positing cognitive mechanisms. As illustrated above, an answer does not need to be based, as the Input-Output model assumes, around citing a process that lies causally upstream of action and downstream of perception. A behaviour, thought, or belief, can be the result of an on-going loop between brain, body, and environment. In these cases, an explanation of how a behaviour, thought, or belief tends to occur involves the whole mechanism. If one chooses to call this wider process not a 'cognitive mechanism', but reserve that expression only for its neural part, that does not alter the fact that our aims in explaining cognition will not be served by a purely internal story. Consequently, restricting the title 'cognitive' to neural activity alone would only achieve a Pyrrhic victory in this context: it preserves the letter of the neural sufficiency assumption, but concedes that it is drained of its power to explain how a cognitive system works. On either view then, consideration of exclusively neural mechanisms does not reveal the mechanisms that explain our cognitive activity.

Not all cognitive processes are environment-involving in the way suggested above. However, I wish to suggest that at least some of the cognitive processes involved in psychiatric disorders are. An embedded model explains how external resources work together with neural activity to produce cognitive and behavioural deficits. It helps to explain why psychiatric illnesses are often tied to interactions with specific environmental props or features of the subject's body. It also helps to explain at least one of the difficulties encountered in fMRI scanning described by Kanaan

[2] See Rupert (2004, 2009) for a sophisticated development of this response, and Sprevak (2010) for a discussion.

and McGuire: trying to illicit the right, characteristically 'ill', behaviour or thoughts from a patient in the alien environment of the scanner. If the cognitive malfunction involved in that the disorder is environment-involving, one would expect difficulties reproducing it inside a scanner.

So what would an embedded model of psychiatry look like?

Consider two hypothetical models of ageing. One is the model of a master clock (or indeed multiple clocks) inside the organism that gradually run down. As the clock advances, the organism ages. A natural thought to prolong life is to somehow slow down, or stop, the internal clock. An alternative model is based on an analogy with how elderly cars age (Hayflick 1999). The thought is that small failures inside a car's mechanism may, by themselves, be untroubling, but these small failures get exacerbated by repeatedly coming into contact both with each other, other components of the car, and an uncooperative environment. The on-going interaction between these elements can grow what initially seem to be small failures in large and unexpected directions, and place them beyond the ability of a repair mechanism to fix. The malfunctioning here does not involve a single localisable internal failure, but is the result of any number of internal abnormalities that are supported by, and reinforced by, the feedback the car receives from its environment.

The fMRI dream of psychiatric imaging follows roughly the master-clock model. Psychiatric disorders are a matter of one (or more) neurological functional failures. The dream is that taxonomising, diagnosing, and treating the disorder can be done in terms of taxonomising, diagnosing, and treating the underlying neurological failures. An embedded account of psychiatric disorders would follow roughly the elderly-car model. Psychiatric disorders may not have a single identifiable internal malfunction. Rather, they are the product of small, perhaps otherwise untroubling, internal misfunctions, that jostling together, and being reinforced by an uncooperative environment, snowball, and contribute to wider breakdown. Just as the failure of an elderly car cannot be pin-pointed to a single localisable failure in a component, so the malfunctioning in a major psychiatric disorder may not be localisable to a given brain region or function. What has gone wrong is that, for any number of reasons, the whole organism-environment loop has been thrown out of kilter, and that can cause, and exacerbate, errors in any portion of the loop. Indeed, it would be reasonable to expect multiple deficiencies both brain-side and environment-side. Curing the problem, even in an ideal case, may not be as simple as achieving correct neural functioning in a given functional brain region. Somehow, the whole organism-environment loop needs to get back on track to function correctly, and this may involve treatment of multiple problems both brain-side and environment-side.

On this view, one would not expect necessarily to identify a single characteristic region or functional neural group co-occurring with a complex psychiatric disorder. All manner of neural malfunctions, when coupled with a recalcitrant environment, may produce similar symptoms. Just as two old cars can have different patterns of internal problems that cause their mechanisms to fail, so two patients may exhibit similar clinical profiles but have different neurological conditions. The embedded model therefore casts doubt not just on the claim that we can reduce the cognitive mechanisms involved in psychiatric disorders to neural mechanisms, but also on the weaker claim that we will necessarily find consistent markers for those disorders within the neural domain. Just as psychiatric disorders generally lack a pathognomonic neuropathology, we may find they also lack a pathognomonic neuro*functional*pathology.

It is worth emphasising that I am not saying that brain malfunction does not play an essential role in mental illness, or that fMRI cannot enhance our understanding the mechanisms of mental illness. My claim is that fMRI may only be part of the story by uncovering the brain-side mechanisms involved in mental illness. The positive proposal is that the mechanism involved in a cognitive malfunction has parts both in the environment and in the brain, and these need to be seen as working in concert in order to effect any kind of mechanistic reduction.

This only sketches how a model of psychiatric disorders might take into account the embedded nature of the mind. The concerns fuel Kanaan and McGuire's conclusion that 'psychiatric fMRI has nowhere firm to plant its feet'. If the concerns raised here are valid, we should not be surprised, or troubled, by this conclusion. On the assumption that psychiatric disorders ineliminably involve not just the brain, but the way in which the brain is coupled to the environment, it should be no surprise that brain scanning is not sufficient to understand their pathology. Failures can occur, not just in a characteristic functional brain area, but anywhere in the loop that extends between the brain and the environment. Once failure in one part of that loop occurs, one should expect more and more brain-side and world-side malfunctions. Major psychiatric disorders, if reducible at all, may only be reducible to a snowballing pattern of internal and external failures.

References

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

Clark, A., and D. J. Chalmers. 1998. 'The extended mind'. *Analysis* 58:7–19.

Dennett, D. C. 1996. *Kinds of Minds.* New York, NY: Basic Books.

Haugeland, J. 1998. 'Mind embodied and embedded'. In *Having Thought: Essays in the Metaphysics of Mind,* edited by J. Haugeland, 207–240. Cambridge, MA: Harvard University Press.

Hayflick, L. 1999. 'Aging and the genome'. *Science* 283:2017.

Hurley, S. 1998. *Consciousness in Action.* Cambridge, MA: Harvard University Press.

Hutchins, E. 1995. *Cognition in the Wild.* Cambridge, MA: MIT Press.

Kanaan, R. A. A., and P. K. McGuire. 2011. 'Conceptual challenges in the neuroimaging of psychiatric disorders'. *Philosophy, Psychiatry and Psychology* 18:323–332.

Kirsh, D. 1995. 'The intelligent use of space'. *Artificial Intelligence* 73:31–68.

McBeath, M. K., D. M. Shaffer and M. K. Kaiser. 1995. 'How baseball outfielders determine where to run to catch fly balls'. *Science* 268:569–573.

Menary, R. 2007. *Cognitive Integration: Attacking the Bounds of Cognition.* New York, NY: Palgrave Macmillan.

Noë, A. 2004. *Action in Perception.* Cambridge, MA: MIT Press.

Rowlands, M. 2003. *Externalism: Putting Mind and World Back Together Again.* Chesham: Acumen.

———. 2006. *Body Language: Representing in Action.* Cambridge, MA: MIT Press.

Rupert, R. D. 2004. 'Challenges to the hypothesis of extended cognition'. *The Journal of Philosophy* 101:389–428.

———. 2009. *Cognitive Systems and the Extended Mind.* Oxford: Oxford University Press.

Sprevak, M. 2009. 'Extended cognition and functionalism'. *The Journal of Philosophy* 106:503–527.

———. 2010. 'Inference to the hypothesis of extended cognition'. *Studies in History and Philosophy of Science* 41:353–362.

Wheeler, M. 2005. *Reconstructing the Cognitive World.* Cambridge, MA: MIT Press.

Wilson, R. A. 2004. *Boundaries of the Mind: The Individual in the Fragile Sciences—Cognition.* Cambridge: Cambridge University Press.