ELSEVIER Stud. Hist. Phil. Biol. & Biomed. Sci. 35 (2004) 219–225

Studies in History and Philosophy of Biological and Biomedical Sciences

www.elsevier.com/locate/shpsc

The brain in a vat

Cathy Gere

Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK

This special issue of *Studies in History and Philosophy of Biological and Biomedical Sciences* has its genesis in an email sent out to a history and philosophy of science discussion list. Researching an obscure topic in the history of the neurosciences, I put out a request for information, with an idle question appended to the end of the message: 'Does anyone know when the brain in a vat thought experiment first appeared in analytical philosophy?' To my surprise, this query triggered a deluge of replies, ranging from the hilarious, to the bizarre, to the gruesome. When it was subsequently suggested that the brain in a vat might provide the theme for a special issue of *Studies*, the proposition was irresistible. I recruited my brother—historian of the digital age Charlie Gere—as co-editor, and we started to solicit contributions, aiming to bring together a broad spectrum of approaches to this eccentric yet oddly fecund theme.

Say the words 'brain in a vat' to an aficionado of science fiction, and she will know immediately what you mean: a disembodied, live human brain, suspended in a vessel filled with bubbling, blood–temperature liquid of some description, experiencing consciousness in the absence of a body. Maybe the brain has been brought out of cryonic suspension to enjoy a form of technoscientific immortality, hooked up to a variety of prosthetic sensory devices giving it access to the world beyond the vat. There will probably be some sort of interface allowing it to communicate its disembodied thoughts, perhaps as a stream of little green letters on a computer monitor, or as a strangely inexpressive voice, speaking, of course, American English. This science fiction vat-brain is an emblem of scientific optimism, a symbol of the unbounded technical progress that may one day allow our species to defy death itself.

Utter the same phrase to an analytical philosopher and she will also immediately recognise the concept: for her, 'the brain in a vat' denotes a thought experiment

E-mail address: cathygere@cantab.net (C. Gere).

1369-8486/$ - see front matter
doi:10.1016/j.shpsc.2004.03.002

constituting a technoscientific update of Cartesian scepticism. The philosophical brain and vat apparatus is roughly similar to the sci-fi setup, but in this scenario an evil scientist is stimulating the tissue with computer-controlled electrodes in order to give the unbodied consciousness all the sensations and appearances of normal embodied life. For philosophers, the envatted brain is a new way of experiencing a very old skepticism. How do *you* know that you are not a brain in a vat? Like Descartes's demon, the malignant neuroscientist in the brain in a vat scenario raises the question of what, if anything, we can securely know about the conditions of our human existence. It expresses the central epistemic worry of dualism, that abyss of uncertainty that yawns in the disjuncture between our inner consciousness and the outer world.

The brain in a vat is an emblem of our technocracy; a vision of scientists as immortality-bestowing gods and illusion-producing devils. These are the enchantments of a disenchanted world: the neuroscientist in the brain in a vat scenario tricks or resurrects not the soul or the mind, but that ultimate emblem of materialist human identity, the intricate, spongy mass of delicate grey tissue at the anterior end of the spinal cord. The envatted brain strips down human subjectivity to the operation of a single organ, and suggests that it may be infinitely malleable by virtue of a sort of ultra-refined digital cattle prod. Our aim in what follows is to situate this symptom of technological hubris in a rich cultural framework, opening it up in unexpected ways, subjecting it not only to the interrogation of philosophers, but also to the scrutiny of artists, literary critics and historians of technology.

One brain in a vat scenario effortlessly dominates the others. If you type the words 'brain in a vat' into an internet search engine, among all the bizarre cultural flotsam (my current favourite being the brain in a vat Lego piece) one name keeps recurring: that of the renowned Harvard philosopher Hilary Putnam. Putnam's famous argument against brain in a vat scepticism in his 1981 *Reason, truth and history* is still, twenty years later, generating responses and counter-responses. In the first chapter of the book he lays out 'a science fiction possibility discussed by philosophers': that your brain has been removed by an evil scientist and hooked up to a 'super-scientific computer' that is stimulating it to produce the 'illusion that everything is perfectly normal'. This scenario, Putnam claims, raises 'the classical problem of scepticism in a modern way' and allows us to probe the 'mind/world relationship' (Putnam, 1981, pp. 5–6). As Mark Sprevak and Christina McLeish explain in their lucid exposition of Putnam's argument—the first essay in this collection—the philosopher's utterly unexpected claim is that the brain in a vat scenario is physically possible but *semantically impossible*.

Throughout *Reason, truth and history*, Putnam's principal target is a philosophical doctrine that goes variously under the name of metaphysical realism, externalism, the correspondence theory of truth, and, more insultingly, the 'God's eye view'. For Putnam, the brain in a vat predicament points towards the impossibility of any sort of God's eye view, any absolutely realistic view from outside the vat that might allow us to say without fear of self-refutation, that we are brains in vats. The same is true of unenvatted knowledge. This move, he grandly declares, signals 'the demise of a theory that lasted for over two thousand years' (1981,

p. 74). For the second essay in this collection we are privileged to have a contribution by one of the most distinguished of the Australian realists who served as Putnam's original foils in *Reason, truth and history*, J. J. C. Smart. Smart's article 'The brain in the vat and the question of metaphysical realism' argues, among other things, that some of Putnam's objections to brain in a vat scepticism might melt away if they can be exposed to the right sort of theory of fiction. Putnam's worry is that the brain in a vat cannot *refer* to anything (least of all its own envatted state). Smart suggests that vat reference is to ordinary reference much as fiction is to history. Just as fiction is pretending to refer to real things (not really referring to non-existent things since there are no non-existent things) so vat reference is rather like a pretend reference to real things.

In the next piece, analytical philosopher Neil Manson asks how Putnam's worries about the radical cognitive disanalogies between real and fictional worlds might affect the claims of a set of cutting edge neuroscientific experiments. A group at Georgia Tech has been studying the neuronal basis of learning and memory with an apparatus in which rat neurons are cultured on a dish covered in a grid of electrodes capable of stimulating and recording neuronal activity. The electrodes are linked to a computer that records outputs from the brain cells and provides inputs to them, in accordance with how a virtual 'rat' negotiates a virtual 'room'. Manson suggests that Putnam's arguments about reference with regard to the brain in a vat can be applied to this experimental paradigm, thus undermining the experimenters' claims that they are able to extrapolate from their envatted rat brains anything about how learning and memory operate in the real world. After rehearsing various solutions to this dilemma, Manson goes on to solve the problem by advocating a *non-representational* epistemology for such experiments, an embodied, situated view of cognition, inspired by the field of Artificial Life, in which the success of an organism in negotiating its environment does not depend upon the mediation of *representations* of that environment, but rather on information processing of a different, less complex, more direct order.

If Neil Manson's model of embodied cognition repudiates the traditional dualism of interior representations of an outside world, the next piece extends this anti dualist stance, delineating a model of cognition in which humans think with their *things*, their tools and their symbols, 'an intelligence not embedded in the head but spread across networks'. Beginning with an unsparing dissection of the rhetorical strategies by which Putnam opens up the abyss of scepticism never, despite his best efforts, to quite close it again, Fred Botting argues that this failure discloses the hollow interior of rationalism, leaving thought 'curiously without core, periphery or limit'. In prose of increasingly ferocious lyrical intensity he leads the reader from Putnam's vat brain into 'Extimatrix', 'an entity formed of the entirety of virtual and digital networks'. The piece is both a vivid evocation and a chilling analysis of the human subject under the conditions of the information age, in which 'the hollows of modernity' are 'excavated and filled by the imperatives and images of hypermodern existence'.

Botting's paper, by invoking the landscape of cyberspace with its attendant sense of paranoia and uncertainty, raises a host of questions and themes that are

developed in the next three articles. Dani Cavallaro's article, 'The brain in a vat in cyberpunk: the persistence of the flesh', interrogates the ubiquity of the brain in a vat theme in the literary and cinematic genre of cyberpunk. Arguing that the 'recurring representation of human beings hooked up to digital matrices' is a central preoccupation through which the genre interrogates the epistemological and ethical issues of technocracy, Cavallaro exposes the sometimes violent or gruesome return of the physical body in all these attempts to picture disembodied thought. The brain, as one of her authors reminds us 'happens to be a meat machine'. The decarnalization of consciousness emerges as a process that can never be completed, continually collapsing under the irreducible materiality of machines and bodies. As one of the participants in an epistemology reading group recently asked about Putnam's scenario, 'What happens when the brain gets damaged?' What, in other words, is the epistemological status of that fleshy interface between the real world and the 'vat-image': the meat machine, hypermodernity's most sophisticated 'wetware', the human brain in all its material delicacy, complexity, and vulnerability?

For Botting, the central paradox of Putnam's brain in a vat is that it never achieves its goal of delivering us from scepticism back to certainty. Cavallaro suggests that any attempt to usher in a regime of disembodiment may end up saddled with more putrid meat than ever. In the next article, John Tresch aims his rhetorical force at another of the contradictions through which the brain in a vat scenario implodes through its own centre. On the one hand, he points out, the all-powerful vat-apparatus implies complete trust in the powers of reason and invention; on the other hand, the scepticism generated by this victory implies complete doubt about our knowledge of the external world. Starting from the 1914 brain in a vat of the proto-surrealist author Raymond Roussel (for which discovery Tresch must be awarded the palm for digging out the earliest example of the trope), he historicizes this paradox, locating it in a tradition of French literature in which *decapitation* features as the site for a highly ambivalent exploration of the politics of rationality versus unreason. The first brain in a vat turns out to be the severed head of a revolutionary rationalism that devours its own makers in a convulsion of barbaric violence.

As these semiotic, literary, and historical analyses show, the paradoxes of the brain in a vat thought experiment lead us into a hypermodern hall of mirrors, an infinite regress of alternating layers of truth and fiction, reality and virtuality. The next piece, an interview with the artist Rod Dickinson, comprises a multi-layered, historically stratified exploration of the very questions of brain-washing, delusion and technical mediation that so beset the envatted. To take but one example, this interview includes a discussion of his film documenting his meticulous four hour re-enactment of the notorious Stanley Milgram 'Obedience to authority' experiments. The Milgram experiments, which took place in the 1960s, asked volunteers to give apparently real electric shocks to other individuals they could observe, under the impression that they were assisting with an investigation into pain. These volunteers were in fact the experiment's subjects, while the apparent subjects of the shocks were actors. The aim was to see if the volunteers would inflict pain on others in the name of obedience to authority. Dickinson's film is thus a recording

of a re-enactment of a series of bizarre performances that themselves comprised a coded re-enactment of another event, the Holocaust, (itself, of course, the subject of elaborate historical denials). In their conversation about this and other works, Charlie Gere and Rod Dickinson reveal the extent to which brain in a vat scepticism is but one small example of the widespread epistemic uncertainties of a media-saturated world.

The brain as a model object through which the place of humanity in the information age can be negotiated and renegotiated is the subject of Charlie Gere's essay 'Brains-in-vats, giant brains and world brains: the brain as metaphor in digital culture'. Gere points out that all the variants on the brain in vat scenario depend for their plausibility upon the notion that the human brain is compatible with man-made technology, and proceeds to narrate a history of modernity as the story of a progressive breakdown in the distinction between the mechanical and the human. He argues that this narrative can be traced through the development of three distinct figures: the 'Brain in a Vat', in which the brain is connected to electronic technologies; the 'Giant Brain', in which the brain's functions can be reproduced by electronic computing technology; and the 'World' or 'Global Brain', in which the connectivity enabled by information-communications technologies produces forms of distributed intelligence. The paper thus contextualises the philosophers' thought experiment, showing how it emerged as just one of a number of tropes in which the flickering of the electrical brain serves as a site for the erosion of the distinction between the human and the technological.

One aspect of the brain's symbolic existence in the digital age is subjected to very close scrutiny in the next essay, Anne Beaulieu's 'From brainbank to database: the informational turn in the study of the brain'. Beaulieu begins by pointing out an analogy between the two principal elements of the brain in a vat scenario—the wet materiality of the brain itself and the digital technology of the computer that regulates its experiences—and two different kinds of neuroscientific collecting activities. The wetware in the vat looks pretty much like a live version of the traditional brain in a jar of the anatomy museum, whereas the computer interface evokes a new kind of brain collection: the brain database. Beaulieu considers what is involved in the shift in type of object, from the scarce, singular, biological brain to the plentiful, digital, replicable one. Contrasting wet and virtual brains as epistemic objects, within experimental systems, Beaulieu shows how the differences between the two types of archive create their own distinctive research agendas.

With the penultimate article, Bronwyn Parry's 'Technologies of immortality: the brain on ice', we finally depart the sceptical nightmares of the envatted brain and embrace, if only for a few heady moments, its dreams of immortality. Beginning with the 1929 futuristic manifesto of John Desmond Bernal, the first properly immortal envatted human brain in literature, Parry narrates a double-stranded history of attempts to preserve and prolong brain function through the freezing of neural tissue. On the one hand she tells the story of the cryonics movement, those much-ridiculed Californians who have had their brains frozen in the hope of technoscientific resurrection. On the other hand, she narrates the development of the highly respectable science of cryopreservation, in which the human body is broken

down into constituent parts and distributed among research programmes and organ recipients. Parry's essay presents a delicate, often hilarious, analysis of the relationship between these two endeavours, illuminating the complicated entanglement between the sciences of prolonging human life and science-fantasies of human immortality.

The last essay in the collection, my 'Thought in a vat: thinking through Annie Cattrell', was inspired by the work of a contemporary artist. Cattrell has been working closely with neuroscientists to produce a set of sculptures in which MRI scans of the different sensory areas in the brain are rendered in three dimensions. These pieces make compellingly visible the doctrine of localisation: the idea that different physical parts of the brain are dedicated to different functions. I realised, as I pondered these objects, that the philosophers' brain in a vat is another outcome of the localizationist paradigm. This realisation enabled me to work out the immediate technological preconditions for the emergence of the brain in a vat thought experiment in the mid-1970s, thus answering the question about its origins with which this whole project began. An interview with Cattrell also revealed that her interest in neural mapping arose from her time spent as artist-in-residence at a psychiatric hospital, lending an urgent moral dimension to her engagement with the neurosciences. Through talking to her I became aware of the work and suffering that made the thought experiment thinkable at all: the brain in a vat is the outcome of a century of experimental medicine, practised on human subjects, mostly epileptics, whose desperation was such that they willingly submitted themselves to untried surgical and technical interventions. It is to the memory of those patients that this volume is dedicated.

My idle question about the origins of the brain in a vat thought experiment turned out to be a surprisingly difficult one to answer. Every so often someone would send me back to Descartes, as though the centuries between the sixteen forties and the nineteen seventies collapsed in the face of the perenniality of philosophical questions. By claiming a place outside history, the envatted sceptic is thus doubly disembodied, shrugging off, in a single gesture, both the messy corporeality of the body and the untidy baggage of time's passing. In my more disgruntled moments I felt like an atheist with an interest in ecclesiastical architecture who wanders into a beautiful church only to find out that no-one can tell her when it was built, so dedicated is the building to the idea of eternity. Faced with a perennial question, I always get snagged rebelliously by the *content* of the philosophical example: the very mention of 'the table in front of us' will send me hurtling off on a train of thought about the non-eternity of tables, the comparative anthropology of furniture design, the history of the disciplining of academic posture.

The most unreconstructed contextualist has to acknowledge, however, that there is a naturalistic version of external world scepticism that has survived so many centuries as to count as perennial: dream scepticism. A Taoist philosopher dreams that he is a butterfly in the fourth century BC and then wonders upon waking if he is not, after all, a butterfly dreaming that he is a Taoist philosopher. Descartes, sitting by the fire in his dressing gown, remembers that he has often dreamt that he is sitting by the fire in his dressing gown and almost persuades himself that the writ-

ing of the First Meditation is itself a dream. Faced with the stability of this worry, demons and demonic neuroscientists begin to look like nothing more than analytically irrelevant variations on a theme so hoary as to defy all historical contingency.

But perhaps I might be permitted to point out that as soon as we progress from questions to answers, the flux of history returns with a vengeance. Chuang Tzu concluded that his flutter of uncertainty between dreaming and waking proved the Taoist doctrine of the transition of all things. Descartes eventually reached certainty not just of his own thinking existence but also that of a majestic male God. Putnam, for his part, finds himself mired in the ubiquitous post-colonial debate between relativism and realism, complete with its usual cast of Nazis and anthropologists, Kuhnians and positivists, and, of course, its historicising tendency. One of the many beauties of *Reason, truth and history* is the suggestion, however lightly sketched, that the philosopher who ignores history will end up as adrift in her search for truth as the historian who repudiates reason.