J. Andrew Ross

The Self: From Soul to Brain

A New York Academy of Sciences Conference, New York City, 26–28 September, 2002

Gurus

The Mount Sinai School of Medicine is an imposing monument to the wealth and power of scientific medicine. Set on its own block in upper Manhattan, its rhetorical centre is the Stern Auditorium. Here, just over a year after 9/11, a group of gurus and self-seekers assembled to confer on the nature of the self. I was there too, looking for help in constructing a grand unified theory of soul and brain.

The New Hubris

Joseph LeDoux, Henry and Lucy Moses Professor of Science at New York University's Center for Neural Sciences, was the man in the middle, the master of ceremonies. It was his idea to bring this event into being, perhaps as a gathering of friends and colleagues to celebrate and confirm their collective status as guardians and cultivators of the new reigning orthodoxy about the self. Not quite incidentally, it also served to showcase his new book, *Synaptic Self* (LeDoux, 2002). From his perspective, the very first words of that book could probably have served as the motto for the conference:

> The bottom-line point of this book is 'You are your synapses'. Synapses are the spaces between brain cells, but are much more. They are the channels of communication between brain cells, and the means by which most of what the brain does is accomplished (*ibid*., p. ix).

At 52, LeDoux stands at the height of his professional standing, well represented by the strength of his new book and his impressively smooth orchestration of the conference. Son of a butcher and set on his course as a teenager by the experience of extracting bullets from cows' brains, he recalls his early research at Louisiana State University thus:

> Robert Thompson was one of the early proponents of the systems approach. . . . Unencumbered by theoretical preconceptions, Thompson marched through the rat

Correspondence: *me@andyross.net*

Journal of Consciousness Studies, **10**, No. 2, 2003, pp. 67–85

> brain, making lesions from front to back and top to bottom, and constructed neural systems of learning and memory totally from empirical observations. . . . I owe my whole career to Bob Thompson (1993, p. 149).

At 52 myself, I have carved a less glorious trail in the course of my pilgrimage to Mount Sinai. My humble contribution in this report will be to do what the invited philosophical gadfly at the conference, Daniel Dennett, in my not-so-humble opinion failed to do, and locate the critical weakness in the synaptic self, so that we can shrug off this new *hubris*. But first, let us recall the events of the conference and relive the excitement of hearing the words of the prophets.

Sessions 1 and 2: Perspectives on the Self

LeDoux opened the proceedings by sketching out the scope of the conference. It was an attempt to think about the self in terms of the brain. What is the self? How does it relate to the brain? Is there room for a soul? How is the self related to consciousness? Is it possible to have more than one self in the same brain? What are the roles of memory, and genes? How does the self relate to personality, and in what sense do other animals have selves?

LeDoux's model of the self is that it is an integrated representational structure distributed over the brain system as a pattern of synaptic connections. The pattern reflects which neurons connect with which others, and how strong the connections are between them. Determine that pattern in a brain and you determine the self that owns or occupies that brain. LeDoux referred to his own earlier work on how synaptic changes caused by stress-induced neural activity centred on the amygdala and hippocampus explain fear conditioning and anxiety states, work reported in his book *The Emotional Brain* (1996), to illustrate how mental states can be reduced in classic scientific fashion to underlying neural activity.

But LeDoux pushes his case. For him, the self is synaptic, period. He can't see what else it could be. He says the trick is to understand how the self emerges from synapses. Indeed. The synaptic story leaves us struggling when we approach the realm of what we used to describe as the soul.

Patricia Smith Churchland was the next speaker. She is professor and chair of philosophy at the University of California, San Diego, and a celebrated public figure. She and her husband Paul have come to symbolize a whole approach to the mind, the hard-AI or *computationalist* approach. She related her talk to her recent paper in *Science* (2002).

In her view, the brain's earliest self-representational capacities arose as evolution found neural network solutions for coordinating and regulating inner-body signals. She sees the neural basis for self-control in natural selection for individuals that were neurally equipped to forego short-term gratification for the sake of long-term reward, and to suppress impulses that had self-destructive consequences. In social animals, this included the ability to modify social behaviour through reward and punishment and to develop skills in cooperative behaviour.

She said that since human brains are very similar to those of other apes and monkeys, the human experience of self is unlikely to be unique. Selves may have evolved

to maintain a basic level of coordination of bodily functions and behaviours (such as the famous four Fs) by using an inner modelling capability to assist in motor planning. The inner model represents the animal's own body in its environment and includes some level of simulation of body, world, self, and other selves.

She asked whether we can give a neural characterization of the contrast between being in and out of control. Whatever the self is, a general formal description may represent it as a multidimensional entity in a large abstract space with dimensions coding information about neuronal organization in various parts of the cortex and in the amygdala and hypothalamus, as well as molecular-level parameters for levels of various neuromodulators, hormones, proteins and so on that influence how we interact dynamically with our environments. A zone within this space represents our being in control, while much of the rest of the space represents our being out of control. If a subject's parametric state puts that person's brain in the 'in control' zone, that person acted freely and is responsible for the relevant actions, whereas a person whose brain is in the 'out of control' zone did not act freely. This is relevant in turn to ideas about the biological basis for ethics and about how best to maintain civil society. At this point in history, the multidimensional space is hand waving, of course, but she pointed out that such models co-evolve with our knowledge of the practical details they are intended to explain. As we learn more, we can refine the idea.

Daniel Schacter was the next speaker. He is professor and chair of psychology at Harvard University, and studies the psychological and biological aspects of human memory and amnesia. His list of publications is vast, and includes *Searching for Memory* (1996) and *The Seven Sins of Memory* (2001), which were among the New York Times Book Review Notable Books of the Year in 1996 and 2001 respectively. His theme was the relation of self and memory as two sides of the same fact, as William James noted in 1890. Our memory of the past is reviewed as a drama in which the self is the leading player. Like all authors of recent books, Schacter reviewed his book. The seven sins of memory are:

- Transience, or decreasing accessibility of information over time,
- Absent-mindedness, or failures at the interface of memory and attention,
- Blocking, or temporary inaccessibility of stored information,
- Misattribution, or assigning a memory to the wrong source,
- Suggestibility, or implanting false memories,
- Bias, or rewriting the past on the basis of current knowledge and beliefs, and
- Persistence, or intrusive recollections that are difficult to forget.

Such 'sins' (I still find this a bizarre word in this context) change the self: citing William James again, my losses of memory or false memories change *me*.

Regarding bias, which is a top-down influence on memory, the most obvious example is the egocentric bias. It seems that information that is relevant to the self is processed in a different frontal region than information that is not. We stabilize our sense of self by seeking to preserve consistency, which involves the sins of bias and misattribution. People tend to misremember past attributes of themselves in line with their present attributes. Our efforts to preserve consistency can be seen as a mechanism to reduce cognitive dissonance with regard to memory.

Schacter's general point — that memories make the self — is entirely consistent with LeDoux's view. When we lay down memories, we change our synaptic connections by growing new ones, pruning old ones, or changing the weights of existing ones. But memory research often relies on first-person techniques such as introspection to find out what a person remembers. With memory, we cannot deny the primacy of phenomenology.

That was it for Thursday. My phenomenology suddenly got wet as, filled with zeal for the new words in my head, I ran 36 blocks south down Madison Avenue through the pouring rain to the conference hotel.

Friday started with a session chaired by Daniel Schacter. The fist speaker was Nancey Murphy, a professor at Fuller Theological Seminary in Pasadena, California. She is a prolific author and helps plan conferences on science and theology sponsored by the Vatican Observatory. Her question: Whatever happened to the soul?

Her main point was that there was no conflict between the emerging neuroscientific consensus and normative Judaeo-Christian views. There could be agreement on the physical nature of humans. She presented a more or less historical survey, from the dualistic distortion of a biblical view that one finds in Hellenistic philosophy — and which reappears in Descartes' dualism — through the neo-Platonic views of St Augustine and the Aristotelian scholasticism of St Thomas Aquinas and others to Kant's transcendental argument for the immortality of the soul. She dwelled sympathetically on the Thomistic view that the soul is the *form* of the body, a view that is held by many modern Catholic theologians. However, if I recall correctly, she said that the crude scientific doctrine of physical atomism made it impossible to regard the soul as the form of the body.

If I may respond here, form is a concept from information theory. To say the soul is the form of the body is to say that the soul is the dynamically evolving configuration of the ultimate parts of the body, and as such is a structure with a mathematical description. The soul therefore enjoys the same eternal quality as any mathematical or informational entity. My soul is coded in a bit string that can be used to call me back into existence for as long as God has the right software. There is nothing here to contradict either the crudest atomism or the most exalted belief in immortality.

Murphy granted gracefully that Christians are free to believe in either physicalism or dualism. She made the useful point that neurobiological determinism did not supplant the concepts of free will and so on, but seemed to require us to develop some new terminology. In sum, she argued that the Jewish and Christian traditions contain minority voices that are not only consistent with the results of current cognitive–neuroscientific research, but also provide grounds for celebrating the monistic–physicalistic accounts of human nature that science promotes.

Alexandre Mauron was next. He is professor of bioethics at the University of Geneva. His research work was in molecular genetics and neurobiology, but since the late 1980s his work has included the ethical issues of genetics and related areas. He began with a reference to the contemporary German-language philosopher Peter Sloterdijk and the provocative idea of the self-engineering of

mankind — *homo faber sui ipsius* — as well as its resonance, in the works of Nietzsche and Heidegger, with the project of domesticating the species. All this clearly needs ethical terms of reference.

In this context, neurobiological visions of shaping the brain seem less controversial than eugenic visions of rebuilding the genome, given that the genome is the ontological hard core of an organism. We can see the genome as the secular equivalent of the soul, and genomic metaphysics as a new kind of hylomorphism. But this suggests that the soul is created at conception, when a new diploid genome is created during fertilization, and this has implications for the ethical standing of the embryo. Yet there are problems with the view of the zygote as a person. What about identical twins or clones — do they share the same soul? Or mosaic individuals — do they have two or more souls? Mauron maintained that a better basic criterion for personhood is numerical identity — one brain, one person.

Another problem with genomic metaphysics is the view it encourages that for any behavioural trait *x* such as alcoholism or dyslexia or homosexuality there is a 'gene for *x*' — that our genes determine our acts. Since genomic characteristics are stable, this view invites an unwelcome fatalism about our prospects for improvement. The idea that our neuronal states determine our acts seems better.

Mauron liked the idea that the self is a social or cultural construct rather than a product of the genome, and in particular liked Sloterdijk's idea that humans create environments — *bubbles* — for themselves that feed back onto human nature and in the long term change us. But he questioned the ethical contrast between genomic manipulation that affects future generations and 'neuromic' manipulation that affects only one individual. As he saw it, *Robocop* and other science-fiction scenarios in which people are transformed by silicon implants create *new* individuals, and hence similar ethical issues. For Mauron, international bioethics is still a very heterogeneous intellectual enterprize, beset with many misunderstandings about implicit standards of argumentation and the proper weight of cultural differences.

Terrence Sejnowski was next. With a Ph.D. in physics from Princeton and an affiliation with Caltech, he is now an investigator with the Howard Hughes Medical Institute and a professor at both the Salk Institute for Biological Studies and the University of California, San Diego. He spoke about the computational self. The brain is never at rest, and its input is a ceaseless pattern of activity. Neurons fire constantly and create a background field, and the task of investigators is to look for patterns above the background. We have learned more about how the central nervous system works in the last ten years than in all of previous history.

Sejnowski considered a hierarchy of scales ordered by powers of ten, from the CNS at the scale of tenths of a meter down to molecules at tenths of a nanometer. This analysis into ever smaller pieces creates the Humpty-Dumpty problem of how to put them all back together again. Microelectrodes can pick up signals from individual neurons, but doing so for a hundred billions neurons at once is impossible. Imagine a brain as big as New York City. Then people in the city are like neurons. Now imagine ten thousand times as many people in the city as there are now, piled miles high into the sky, all communicating with each other busily.

That's an image of the brain. How can you get a meaningful picture of what's going on by tapping the signals from a handful of people?

One way out is to take EEG readings of the total signal. Now we face the cocktail party problem. How do we extract individual signals from the background noise? Sejnowski has developed a computer algorithm to do just that. He calls it the *brain microscope* and sees it as heralding a new dawn of imaging studies. The idea is to record signals from different directions and then perform a *principal component analysis* and an *independent component analysis*. By analysing enough signals, he can generate clear diagrams of what's going on where in the brain. He can see structure in event-related potentials, which consist of large numbers of microvolt-level signals that are averaged out but which also show systematic phase shifts and increased coherence compared to the background. He showed us some fascinating detailed studies. It all seemed very geeky, with huge but vague promise for the future.

A panel discussion followed. In reply to a long question that involved the assertion that realism and idealism are equivalent, Sejnowski said 'one man's top is another man's bottom'. To a question about the chimp in the mirror and whether self-awareness was a test for the existence of a self, he replied that he once wrote a 450-page book about falling asleep in which he discussed the low-frequency, high-amplitude synchronous waves generated by the neurons in that state. He contrasted the waves he studied with the higher-frequency gamma rhythms studied by Wolf Singer and his colleagues. Sejnowski was interested in how thalamo-cortical loops recruit neurons and get them to burst in synchrony. He said he was still trying to put Humpty-Dumpty back together again.

To another question, Professor Murphy said we're all in the process of recognising the falsity of the reductionist assumption that it's all ultimately physics. She opined that we haven't begun to think about complex systems in ways that show *how* that's false. We need to find a way to describe how we become *creators of ourselves*. That brought me back to Nietzsche and Heidegger but left me sceptical about her physics. She's writing a book about it.

Sessions 3 and 4: Psycho-social Aspects of Self

After a kosher lunch in the Mount Sinai canteen, I was ready for the afternoon session chaired by Professor Churchland.

Marc Hauser was first, on 'our ancient selves'. Hauser is currently a professor at Harvard University and author of over 100 peer-reviewed publications. His latest book, *Wild Minds* (2000), is being translated into seven languages. His research sits at the interface between evolutionary biology and cognitive neuroscience and is aimed at understanding the processes and consequences of cognitive evolution. In his talk, he explored how human and nonhuman animals differ with respect to their sense of self. In the first part, he explored the general problem of what animals know about the physical world and revealed an intriguing dissociation between perception and action. In the second part, he examined the capacity of animals to imitate, recognize their image in a mirror, and represent the beliefs and desires of others.

Hauser reminded us that we share 98 percent of our genes with chimps and raised a laugh with a portrait of a chimp morphing into President George W. Bush. He presented some results of his recent research on delayed gratification, reciprocation, and defection, comparing and contrasting human infants and monkeys. One set of experiments involved falling balls. If an experimental subject sees a ball falling toward a table, and then sees a box on the table and a box under the table, which box does the subject approach to try to find the ball? Monkeys, it seems, don't really understand tables and approach the lower box. Human infants do the same. Both human infants and adult monkeys show evidence of exquisite object knowledge, yet appear incapable of accessing such knowledge for the purpose of explicit action. Maybe they have difficulty accessing it because they have weak inhibitory mechanisms, which causes them to engage in ballistic action sequences. There is a gap between perception and action. We can understand a physical regularity yet still act as if we didn't.

As for delayed gratification, Hauser reported some longitudinal studies that tracked infants into adulthood. Long ago, some infants were faced with one sweet now, unconditionally, or two later, if they could first resist the offered sweet for a few minutes. Those who were unable to resist the temptation turned out later in life to have higher rates of alcoholism, gambling, drug abuse and the like, as well as lower SAT scores, job satisfaction and so on, than the stronger-willed infants. This could be relevant to social policy — test infants and plan their lives accordingly.

Regarding altruism, the evolutionary story involves similarity of genes. We are more altruistic toward people who share more genes with us or are more similar to us. In prisoner's dilemma experiments, where cooperation between pairs is rewarded only if both play along, the reciprocal exchange involved is limited by individual strength in the face of delayed gratification and temptation to defect. In experiments with paired monkeys, cooperation continues until one defects, then the wronged partner punishes the defector for a while to produce renewed cooperation. But monkeys can be remarkably altruistic: they will starve themselves rather than administer a painful shock to a fellow, regardless of the dominance relation between the pair. This makes them ethically better than some humans, who will happily shock their fellows if an authority figure tells them it's OK.

Hauser's general message was that our actions lag behind our knowledge in the sense that our actions are in part hard-wired by our evolutionary past and robust against quick change. He suggested that we look to the neural circuitry underlying inhibition and conflict monitoring for clues to the evolution of a human sense of self.

Naomi Quinn was next. She is a professor at Duke University. To quote her conference abstract, she is part of a current effort in cognitive anthropology to build a theory of culture on the basis of schema theory and connectionist modelling, and within this framework to demonstrate how meanings become internalized, shared, motivating, enduring historically and within individuals, and thematic across cultural domains. Whew! Unfortunately, she had no slides to

show us and mumbled indistinctly as she peered myopically at her notes. But I shall try to do her justice.

She reviewed some recent cross-cultural studies of child development, which showed that all cultural models for child rearing work in two ways. First, they all promote extreme constancy of the child's experience, as this is seen to relate to key values and associated behaviours. This constancy is achieved by maintaining a community of opinion about what children must learn, by investing this opinion with moral force, and by embedding it in child-rearing practices that are highly regular and oft repeated. Secondly, the models couple the lessons with techniques to make the learning experience emotionally arousing. In other words, child rearing everywhere is designed to ensure that children get the message, and that they remember it once they get it.

She gave us a handout with some fascinating quotations from various anthropological sources that contrasted child rearing practices in different societies around the world. They were fine anecdotes, but what did they tell us about the self? Professor Quinn: what results from the experience of being reared according to a given cultural model is a lifelong self that is culturally distinctive. Child rearing is the central way to form a self. Quinn concludes that there is no way to say what's the best way to raise a child. And what about the adult consequences of these practices? Or their relevance to neuroscience? Here Quinn squinted and mumbled a *bon mot* — 'I'm waiting for the neuroscientists to fetch me!'

Michael Lewis was next. His theme was the emergence of consciousness and its impact on children's development. Lewis is Director of the Institute for the Study of Child Development at Robert Wood Johnson Medical School, and a professor at Rutgers University. His forthcoming book, *Altering Fate: Why the Past Does Not Predict the Future*, argues that children's conscious adaptation to the current environment is the directing force in development. In his talk, he explored how the adult machinery of the self and the mental state of the idea of *me* develop over the first two years of a child's life, and provide the scaffolding for the child's further development in the social, emotional, and cognitive domains.

He started by considering imitation in newborn infants as a process of sensorimotor integration. Then he considered self-recognition in mirrors, the use of personal pronouns, and pretend play, all of which show that there is a strong developmental coherence in the emergence and onset of an explicit self. They lead to a change in the child's emotional life that is better characterized by embarrassment than self-consciousness. He mentioned his earlier book *Shame: The Exposed Self* (1992). As the child develops the mental state of me, infant features are transformed into humanlike abilities. Social interactions become relationships, primary emotions become self-conscious emotions, and the child develops a theory of mind. He closed with some videos of small children showing how self-recognition and the ability to distinguish between appearance and reality emerge at various mental ages.

Following refreshments, Michael Gazzaniga chaired part two of the session. Gazzaniga is a giant figure in neuroscience. Currently a professor at Dartmouth College and Director of Dartmouth's Center for Cognitive Neuroscience, he is

well known for his deep involvement in matters of public policy, ethics, and public understanding of science.

He introduced Hazel Rose Markus. She is currently a professor at Stanford University. Her research has focused on the role of self in regulating behaviour, and her most recent work explores the interdependence between psychological structures and processes and sociocultural environments. Interestingly, for me at least, she is a member of the MacArthur Research Network on Successful Midlife Development — maybe she would give us a handy tip or two!

Her talk was on models of agency, and her main theme was the contrast between two different ideals for the self in society, symbolized by the metaphor of a fish that can either swim against the stream or just go with the flow. She characterized these as the disjoint and conjoint selves, respectively. The most extreme exaltation of the disjoint self occurs in contemporary U.S. American society, where the independent, self-sufficient individualist is celebrated above all. Contrariwise, the most extreme celebration of conjoint selfhood occurs in China, Korea, and Japan, where interdependence is stressed and they say the nail that sticks out gets hammered.

Disjoint and conjoint selves reflect two contrasting models of action. For the disjoint self, a good act is self-focused and independent — ideal Americans think and act for themselves — with the result that differences between people are affirmed and celebrated. For the conjoint self, a good act is focused on relations with others and on their welfare. Such acts affirm the community and arise from respect or concern for others. Markus illustrated this contrast with slides presenting a barrage of media images, mostly commercial advertising for everyday consumer products. Their overwhelming endorsement of the transpacific contrast was more than just amusing, it was startling. Advertising exploits existing cultural ideals, of course, and can only work when the ideals of self are there to be exploited, but the relentless assault of the media on consumers also reinforces and exaggerates these stereotypes to absurd extremes.

Markus emphasized that these contrasting models of agency are not just in the head but are played out in the world. She reminded us of George W. Bush's call to defend the individual against the collective, as if we were the crew of Starship Enterprise fighting the Borg. She quoted a Japanese person as saying, 'I behave in order for people to feel peaceful'. She told a story about some Korean Americans who visited South Korea and said the visit changed their souls. She made the point that different social contexts shape the individual self to become either conjoint or disjoint. As a European who selfishly conjoins the transpacific disjunction, I can only agree with her.

Daniel Wegner was next. He is currently a professor at Harvard University. His work is focused on the role of thought in self-control and in social life. His anthology *White Bears and Other Unwanted Thoughts* (1989) is fascinating. The title essay explores the psychology of the classic conundrum of trying *not* to think about a white bear, where the harder you try, the harder it gets. His latest book, *The Illusion of Conscious Will* (2002), was the source text for much of his talk. Not to beat about the bush, here is the key to that book, from its preface:

> Do we consciously cause what we do, or do our actions happen to us? . . . This is a book about a different sort of answer to the question. Here it is: Yes, we feel that we consciously cause what we do; and yes, our actions happen to us. Rather than opposites, conscious will and psychological determinism can be friends. Such friendship comes from realizing that the feeling of conscious will is created by the mind and brain just as human actions themselves are created by the mind and brain (p. ix).

In his talk, Wegner addressed the question of why we feel that we cause our actions. He showed a fine slide to get us in the mood — *The Mind's Self-Portrait* by the Dutch engraver M.C. Escher, which portrays a hand holding a mirrored ball in which the image of Escher looking at his image is reflected. As Wegner put it, we see our selves or our souls each time the mind looks at itself. He pursued the metaphor of a painted self-portrait. Our self-image consists entirely of conscious phenomena, so the palette is limited, with no dark shades. The portrait is a miniature, to fit into the mind's tiny space (speak for yourself, Wegner). It presents a model focussed on agent causation, not event causation. And the image is somehow convincing or self-luminous. In our self-portrait, thought seems to cause action, yet in fact our thoughts are part of a much more complex picture.

Wegner was attracted by David Hume's notion of will as the sentiment or feeling we get when we do something. We have the feeling that we cause what we do, but as Hume famously insisted, causation is just constant conjunction, so the way is open to declare that the feeling is an illusion. Experiments tend to support the idea that we rely implicitly on three principles to decide when we willed our actions:

- Consistency: when a thought is relevant to and compatible with the subsequent action, we tend to think the thought caused the action.
- Exclusivity: if there is no other cause for the action in sight, we are free to think we caused the action.
- Priority: the thought must precede the action by an appropriately brief interval.

Wegner discussed various middle-class pastimes from a hundred or more years ago, when people liked to indulge in table-turning, automatic writing, hypnosis, divination and so on. In all these cases, the sense of agency is somehow effaced or diluted, or an 'agentic shift' occurs. A modern analogue of these pastimes is facilitated communication, where a communication-impaired individual enjoys the help of a facilitator to enter text on a keyboard, perhaps by letting the facilitator hold their hand as it twitches over the keyboard, or by letting the facilitator complete sentences or expand on themes. Wegner cited detailed studies of fairly obvious facilitator interference, even when the facilitators were convinced they were not corrupting the messages. His recent book goes into much more detail — some of it very amusing and recounted with admirable wit — on all these activities and how they fit the three principles of agency.

His big conclusion is that whenever we think we willed our action, the brain provides both the thought and the action. The best way to see it is that on the basis of our feelings we *theorize* that we will our actions. Like all theories, the theory of free will is fallible, and may be plain wrong.

Mahzarin Rustum Banaji came on next. Banaji is currently a professor at Harvard University and at Radcliffe. She studies human thinking and feeling as it unfolds in social context. She is particularly interested in the unconscious nature of assessments of self and other humans that reflect feelings and knowledge about social group membership — about age, race or ethnicity, gender, class, and so on. I noticed that she is a young woman of colour.

Her talk was about the unconscious and social construction of preferences and beliefs. She cited some detailed studies of ethnocentrism involving thoughts or feelings about black and white, poor and rich, foreign and American, Jewish and Christian, gay and straight, and trees and birds (the neutral control), where their relation to good and bad was explored. The biases of the experimental subjects were measured in terms of reaction times, and were of course both strong and strongly correlated to the subjects' own positions in all these pairings. Moreover, in each case the majority group showed the bias more strongly.

She then discussed implicit cognition and implicit attitudes that we are unable to identify introspectively. In particular, we have a self-attitude that involves investing objects with associations to our own self. She discussed a study of women versus men in mathematics and science that showed an interesting dissonance between explicit and implicit attitudes for female scientists, who showed little or no explicit bias in thought or feeling against female scientists yet revealed some implicit bias, presumably reflecting the culture around them.

That was the hard work over for the day. We all relaxed at a wine and cheese reception in the lobby. I talked with two nice young ladies and then with a Floridan called Gordon Johns who's writing a book called *The Mythical Me.*

Sessions 5 and 6: Self and Brain

Saturday dawned sunny. Perhaps the shining truth would be revealed at last. The first session was chaired by Joseph LeDoux.

The first speaker was Francesca Happé, who after research on autism under Uta Frith at Oxford is now a senior scientist at the Institute of Psychiatry, King's College London. She told us about some recent research on 'theory of mind' and the self, where theory of mind is the aspect of social cognition that enables us to attribute mental states such as beliefs and desires to others. Despite much research on autism, which is a developmental disorder of social insight, little research has addressed the normal and abnormal development of insight into one's own mental states.

Each of us has a theory of mind. We use it every day to deceive, joke, teach, gossip and so on. Experiments on the development of this theory in children may involve, for example, the famous Smarties task where a tube ostensibly containing chocolate candies but in fact containing something else, such as pencils, is traded between knowing and unknowing kids to see how well they cope with the deceptions involved. The conventional view is that we need a theory of mind for others and that we have privileged access to our own mental states, so we don't need it for ourselves. But Happé reported results that show otherwise. It turns out

that children are no better at attributing mental states to themselves than to others. And autistic children also have problems reading their own minds. It seems that we theorize our own states of mind no less than those of others.

Question: do we activate the same brain regions to read our own and other minds? Neuroimaging studies show that theory of mind activity occurs in medial frontal cortex and paracingulate cortex for both kinds of mind reading. And in both cases, autistic subjects show decreased paracingulate activation in theory of mind tasks compared to normal subjects. So we seem to use similar resources for reading our own and other minds. More speculatively, our ability to read other minds may even precede and facilitate our ability to introspect. Evolution may have forced us to read other minds before our own.

Antonio Damasio was next. He is a professor at both the University of Iowa and the Salk Institute, and has received countless distinctions and prizes, including the Golden Brain Award in 1995. His books *Descartes Error: Emotion, Reason and the Human Brain* (1994) and *The Feeling of What Happens: Body, Emotion, and the Making of Consciousness* (1999) are taught in universities worldwide.

He talked about feeling and self. He distinguishes two kinds of self: *core self* and *extended self*. Core self corresponds to the transient process that is continually generated relative to any object with which an organism interacts, and during which a transient sense of knowing is automatically generated. It requires neither language nor working memory, just short-term memory. Extended self is a more complex process that depends on the gradual build-up of autobiographical memory. It requires conventional memory and is enhanced by language.

Damasio said the essence of the self was its stability, continuity, and singularity. The self is a stable representation of individual continuity and serves as the reference for mental states. Its basis is the representation of one's own body. He finds support for this conception in the writings of Spinoza, William James, Nietzsche, Husserl, and Merleau-Ponty. The representation of the body is the backbone, so to speak, of the representation of self. Its variance has a narrow range, in contrast to the variance of perception, which can approach infinity.

The senses are relevant to the self, but these are not just smell, taste, touch, hearing, and vision. Kinaesthesia and visceral input are also important. Sherrington produced a classification of the senses that distinguished chemoreception, proprioreception, exteroreception, and telereception. The chemoreception system in particular provides a rather detailed sensory representation of the state of the organism. All the input to the brain about the body is much like external sensory input, except that it is steady. The brain uses it to create an *image* of the self.

There I can let it go. Here was a master at work. I look forward joyfully to his forthcoming book, *Looking for Spinoza: Joy, Sorrow, and the Human Brain*.

Rodolfo Llinás was next. Currently a professor at the New York University School of Medicine, his honours include the UNESCO Albert Einstein Gold Medal Award in Science and election to the National Academy of Sciences. His book *I of the Vortex: From Neurons to Self* (2001) introduced the *mindness state* as the class of all functional brain states in which sensorimotor images, including self-awareness, are generated.

He talked on cognition as a premotor event. Consciousness of self is required in order that we can move with intentionality. He used the image of a tennis player to argue that prediction is essential to skilled movement. To laughter (for he is a *maestro* of precise timing and wording), he said that without prediction we would be like bureaucrats, who use their wits to become sessile and then eat their brains because they don't need them any more (he didn't say professors with tenure).

A key event in the biological history of the self was the evolution of neurons that were neither sensory nor motor but *interneural*. Now arbitrarily complex interneural circuitry could develop. In Llinás' view, the central architectural feature here is that of thalamo-cortical loops. The thalamus is at the centre of connectivity to the cortex, and the cortex reconnects to the thalamus 'with a vengeance'. Consciousness, on this view, is a *process*, not a thing, generated by recursive looping between thalamus and cortex.

He talked wide-eyed about cells with 'personality' and 'points of view' and mentioned polarized and depolarized states of the thalamus. Roughly speaking, polarized means you're 'on' and depolarized means you're 'off'. When there was no electricity flowing, there was no *you*. The gamma-band activity of the cells is related to cognition and consciousness. He showed some slides depicting cortical activation in the 35–45 Hz band and showed some MEE images of such activation for dreaming subjects. Dreaming, he said, is similar to wakefulness without the sensory input. The brain is about making images — we could call it a *dreaming machine* limited only by its sensory input.

The thalamo-cortical loop mechanism creates a set of global oscillations that cause the brain to operate discontinuously, generating dreams with a rhythm, like the frames of a movie. New sensory input is fed into this cycle very selectively, with the result that we can focus consciously on only a few things at once. All this was music to my ears — but it was soon over.

A questioner asked why, if the thalamus is so important to the self and we have two thalami, we nevertheless have only one self. The maestro replied simply that we also have a corpus callosum, and thus segued to the next speaker.

Antonio Damasio chaired the final, climactic session, and Michael Gazzaniga was the first of the trinity of climactic speakers. Gazzaniga did his early research under Roger Sperry, who won the Nobel Prize for split-brain research on patients whose corpus callosum had been cut, and this talk was essentially an update on the topic.

To summarize, dividing the cerebral hemispheres of the human brain creates two largely independent cerebral processing centres, each with its own set of mental capacities. The left hemisphere is heavily committed to rational and interpretive functions and the right hemisphere is specialized for visuo-spatial and complex perceptual processes. Observing one's own behaviour creates the subjective sense that a self-directed cognitive system is in action. Gazzaniga and his colleagues recently showed that the left hemisphere has a greater sense of personal self than the right hemisphere.

He started his talk with his conclusion, just in case he ran out of time: *all reality is virtual*. He showed us a slide with a view from above of a split-brain patient

looking at a screen that was divided so that each eye saw just half the screen. The patient was shown images, left and right, and was asked to press the most appropriate keyboard images, again left and right, with the corresponding hand. For example, a screen image may be snow and the keyboard image a shovel, or the screen image a chicken and the keyboard image an egg. When the right eye saw a chicken and the left eye saw snow, the subject was asked why the left hand chose the shovel. 'To shovel up the chicken droppings', came the instant reply, which shows how skilled the subject's verbal hemisphere was at confabulating to cover up its ignorance at what the right hemisphere was doing, as well as how naturally the left hemisphere took the leading role.

It seems that the thinking left hemisphere (depicted as the Einstein in another slide) is good at detecting *self*, while the reactive right hemisphere (depicted as the rat) is good at detecting *other*. The left hemisphere asks how relevant new input is to *me*, the right asks what orientation it has and so on. So the mapping of self into the left hemisphere is natural.

Split-brain patients show no particular insight into their condition. This absence of awareness is typical in cases of brain damage. A subject who suffers a lesion to the visual system may typically complain, 'I can't see any more!' But a subject who suffers a lesion to the visual cortex may not even notice the deficit. The faculty is just gone. The self in the post-lesional brain will regularly interpret the bizarre as normal. To a ripple of laughter, he pointed out that we all do much the same: we interpret a few fancy colours in a VR display as bizarre but treat flying five miles high in an aluminium tube as perfectly normal. He mentioned four syndromes that share this lack of insight:

- Anosognosia, or unawareness of a neurological disability, the most common form of which is unawareness of paralysis,
- Capgras syndrome, in which patients make delusional misidentifications of people they should know, often claiming that the misidentified person is an impostor or double of the 'real' person,
- Reduplicative paramnaesia, or the mistaken belief that there are two nearly identical versions of a particular place, and
- Hemispatial neglect, in which a patient ignores stimuli on the side of the body or the space opposite to a brain lesion.

All these syndromes are described via case studies in *Altered Egos* by Todd Feinberg (2001).

Subjects with these syndromes can show an amazing lack of insight, which must have implications for the insight the rest of us enjoy. Each of us has a self for which the processing is distributed over two hemispheres but which is coordinated into a single, seamless consciousness. Our speech engine does duty for both hemispheres and is ready with a story whether it knows the facts or not. Our consciousness seems integrated on the surface, but the idea that everything comes together in the Cartesian theatre is an illusion. Gazzaniga could have used this line to hand over to Dan Dennett, but Dan had to wait.

Eric Kandel was the second member of the ultimate trinity. He is quite old now, but still spry. He is a professor at Columbia, Senior Investigator at the

Howard Hughes Medical Institute, and the recipient of countless distinctions and awards, including the Nobel Prize in Physiology or Medicine for his work on the molecular mechanisms of memory. His talk was entitled 'Radical Reductionism in Science and Art: Biology of Memory Storage and Minimalist Art'.

Kandel finds memory and learning endlessly fascinating, and is delighted by the idea that new imaging techniques may one day enables doctors to say, for example, 'Here, your superego is a little too large'. We all laughed, and he allowed that this may be less a fond hope than a fond illusion. He discussed the contrast between the blank slate view of the mind and the Kantian view that we come into the world with a toolbox of *a priori* concepts. Some brain science was needed to decide between these views. When he was young, Kandel expected that the black box brain would open up in his lifetime, and indeed he saw the shift from a psychoanalytic view of the mind to empirical biology. For example, in memory, the contrast between explicit or declarative memory and implicit or procedural memory is now explained biologically. Explicit memory involves the temporal lobes and the hippocampus, whereas implicit memory involves the striatum, amygdala, cerebellum, and reflex paths.

Kandel began his work in 1957/58. He realized that the complexity of memory systems was fairly intractable with prevailing techniques and that he had to simplify. So he chose to study the marine snail. This humble organism has about twenty thousand neurons, compared to the human trillion or so. When subjected to stimulus–response training, the marine snail only adapts a few hundred cells, so he could trace the pathways exactly. His research corroborated the Kantian picture. The *patterns* of neural connection were given, hard-wired, and all that changed during learning were the *strengths* of the synaptic connections.

However, the full picture was subtler. Kandel's first stab here was like minimalist art, like a Matisse canvas (pause to show a nice slide). He had to go further. Long-term memory requires the activation of genes. This activation causes the cells to grow new synapses. There are also repressors to prevent all learning from generating long-term memories, so there was a high threshold for such new growth. Still, a new conclusion emerged: your experience affects the expression of your genes. In the long term, your memories are the result of anatomical changes to your brain. The sensorimotor homunculus is not fixed. We all have different brain maps.

Kandel insisted passionately that such biological reduction does not trivialize or reduce the wonder of these natural phenomena. He compared this to Rothko's minimalist art (here we saw some more nice slides), with its move from figurative representation through cubism and the abandonment of form to bands of colour, and finally to sheer black. Rothko saw all this as expressing basic human emotions. Kandel was interested in the power of such reduction. Why do we respond to it as we do? Here he referred to V.S. Ramachandran, who explains it in terms of a limit to our attentional response. We prefer simple images because they enable us to focus on the essentials. This intensifies the pleasure they generate and gives us the kick we crave. As Kandel sees it, art can teach us about how the brain works.

Daniel C. Dennett was the final, ultimate speaker. Billed originally to talk on Friday, he delayed his arrival until Saturday, with the result that his imminent Coming was announced several times during the proceedings. Modestly listed in the program as University Professor, Austin B. Fletcher Professor of Philosophy, and Director of the Center for Cognitive Studies at Tufts University, Dan Dennett is one of the greatest living philosophers and a prolific author, with numerous books and over 200 scholarly articles to his name. His books *Consciousness Explained* (1991) and *Darwin's Dangerous Idea* (1995) are wonderful. Moreover, together with wayward genius Douglas Hofstadter, he edited *The Mind's I — Fantasies and Reflections on Self and Soul* (1981), which is surely one of the best collections on the theme ever assembled.

Dan Dennett is a big man with a silver beard and the charismatic presence of Santa Claus or the Grand Oral Disseminator of Maxi Jazz fame. He went straight to work. Descartes identified *res cogitans* with the immortal soul, but since then materialism has swept dualism aside. We now have a mortal and material soul whose only Cartesian relic is the Cartesian theatre. All the work of the soul is now done by Cartesian theatre homunculi, and all this work must be distributed to lesser agencies in the brain. With the self as an organ, we face Jerry Fodor's big question: Who's in charge?

Dennett's oratorical flow was so fast and rich I that could hardly keep up with my scribbled notes. I wrote: 'Will Hamilton's question — What did I want?' Presumably this was Hamilton the evolutionary biologist. But what did he want? I can venture the brave guess that the point is that such a question is in principle unanswerable because there is a failure of reference in the prerequisite attribution of determinate desires to an ill-defined self.

Or what about his reference to Robert Wright's book *Nonzero: The Logic of Human Destiny* (2000), where in chapter 21, footnote 14, Wright says he's convinced Dennett thinks consciousness doesn't exist? What can we do with this? Dennett's own argument in *Consciousness Explained* was that the heterophenomenology of consciousness reduces it to computation and behaviour, or as John Searle said, explains it away, so why should we insist that Wright is wrong?

At least this was clear. Dennett strongly recommended that we read a book called *Breakdown of Will* by George Ainslie (2001). Apparently, Ainslie argues that an organ of self doesn't have to exist.

Dennett strode on. Why does it *seem* to us that there's a Cartesian theatre? And who is the *us* to whom it so seems? Consider Dan Wegner's claim that we inhabit an extremely complex machine. Who are *we*? Or consider one of Libet's famous experiments: a subject is seated in front of a clock face with a dot on it that rotates around the face at the rate of three and a half revolutions per second. The subject is asked to flick a finger — *flick!* — voluntarily, at whim, and note the position of the rotating dot when the urge to flick emerged in consciousness. Libet's measurements showed that the readiness potential grew in the subject's brain a full 350 milliseconds before the urge emerged. What are we to make of this? In *Consciousness Explained*, Dennett said that although such results seemed to show we were not quite out of the loop, this whole picture was 'compelling but incoherent' (p. 164).

Ramachandran (him again) said (in his 1998 book) that such delays show we don't have free will. What we have is 'not free will but free won't'. That is, the role of the will is inhibitory. In most cases, we just do what we do, but occasionally we can stop ourselves.

Earlier, in his book *Elbow Room* (1984), Dennett said that if you make yourself really small, you can externalize everything. But wherever you are, you get illusions of simultaneity. He showed us a few diagrams to clarify this. Imagine various brain modules communicating with each other. Wherever the self is located relative to these modules, there are varying distances to the respective modules. Given slow signal propagation, this means varying delays caused by the signals' travel times. I can have this signal arrive before that one, or vice versa, by locating the self nearer to this or that source module. But wherever I put myself, the time ordering generated by the arrival of the signals will in general be different from the order of creation of the signals in the modules. That's just physics.

Dennett now suggested various hypotheses. First, the *strolling you* moves back and forth in about 300 milliseconds, so you misjudge all the times. Second, the *out-of-touch you* outsources or delegates the whole business and is thus in a poor position to judge any timings, since all your information is second-hand. Third, you go for Libet's window of opportunity, which looks horribly like an artefact of the whole experimental protocol.

He suggested a new approach. Whenever you distribute work in time and space, you distribute responsibility in time and space, too. In this case, you're not so much out of the loop as the loop itself, the whole thing. That sounded much better.

What was the punch line of all this? Aha, he said, see my next book!

From Ground Zero to Paradise

That was it. A first-class portrait of the current state of play with regard to the neurological self. The big question is whether this snapshot makes it plausible to suppose that the neurological self can do duty for the soul. Or is Joe LeDoux expecting too much of neurological reductionism?

The problem is that the radical reduction this mechanistic metaphor has accomplished leaves us at Ground Zero, with no easy way back to the dizzy heights we wanted to explore. Imagine a tribe of truth-seekers who knew nothing of computers but who were full of hypotheses to explain how they worked. A tribal guru who said a computer was just a pattern of connections between transistors would rightly be celebrated, but could hardly be said to have told the whole story, any more than the earlier guru who said it was just a big pile of elaborately juxtaposed atoms. LeDoux is a synapse specialist, so he sees the brain that way, but the holistic problem of how it all comes together remains untouched.

The neurological self is but one slice or aspect of a many-splendoured thing. The self of popular discourse is so polymorphous that no tidy definition can wrap it up. We have a personal self, a rational self, a conscious self, a biological self, a genetic self, an immunological self, and now a neurological self. Are they all identical? That seems impossible. Each of these selves is defined in a different realm of discourse, and the discourses do not admit straightforward translation

from one to another. Or, to placate Dan Dennett, each is a draft in a multiple-draft drama stretching across our whole civilization. We have a turbulent pandemonium of selves jostling for supremacy in a public theatre. The hope that the whole riotous show boils down to synapses is hollow, not inspiring.

Indeed, LeDoux's very first words in chapter one of his new book give us pause for doubt: '"I don't know, so maybe I'm not", the T-shirt said' (2002, p. 1). This post-Cartesian *aperçu* is my point of critical departure. It hints that the self is the referent of the word 'I' and therefore as polymorphous as our usage of that little word is multifarious. Ever since Moses heard the great I AM, the first-person singular pronoun has been a battleground. The history of 'I' is most unlikely to reduce without residue to talk of synapses. For a conscious mind rises much higher above its synapses than a computer above its transistors. As Llinás said, the electricity has to flow for us to be 'on', and as McFadden says (2001; 2002), the brain's electromagnetic field may be the real substrate for an integrated self. The field generated by billions of neurons reacts back on individual neurons and interacts with the internal and external environment in a dynamical coupling that physicists are still exploring. Our thoughts do not reduce as neatly to synaptic action as computer programs reduce to transistor action.

The cerebral EM field is still *terra incognita*. This is the critical weakness in the neurological concept of self. Perhaps the photonic self will one day be seen to rise as far above the neurological self as the neurological self rises above the genomic self. Perhaps we shall even glimpse a hierarchy of selves, soaring through the hierarchy of Buddhas into Cantor's transfinite paradise.

Buy

In my humble opinion, the conference was a big success. The New York Academy of Sciences is publishing the proceedings in its Annals series. You can buy a copy via www.nyas.org.

References

Ainslie, George (2001), *Breakdown of Will* (Cambridge: Cambridge University Press).

Churchland, Patricia Smith (2002), 'Self-Representation in Nervous Systems', *Science*, **296**, pp. 308–10.

Damasio, Antonio R. (1994), *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Grosset/Putnam).

Damasio, Antonio R. (1999), *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (New York: Harcourt Brace and Co.).

Dennett, Daniel C. (1984), *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press).

Dennett, Daniel C. (1991), *Consciousness Explained* (New York: Little, Brown and Co.).

Dennett, Daniel C. (1995), *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (New York: Simon and Schuster).

Feinberg, Todd E. (2001), *Altered Egos: How the Brain Creates the Self* (New York: Oxford University Press).

Hauser, Marc D. (2000), *Wild Minds: What Animals Really Think* (London: Allen Lane).

Hofstadter, Douglas R. and Dennett, Daniel C. eds. (1981), *The Mind's I: Fantasies and Reflections on Self and the Soul* (New York: Basic Books).

LeDoux, Joseph E. (1993), 'Emotional Memory: In Search of Systems and Synapses', *Annals of the New York Academy of Sciences*, **702**, pp. 149–57.

LeDoux, Joseph E. (1996), *The Emotional Brain: The Mysterious Underpinnings of Emotional Life* (New York: Simon and Schuster).
LeDoux, Joseph E. (2002), *Synaptic Self: How Our Brains Become Who We Are* (New York: Viking Penguin).
Lewis, Michael (1992), *Shame: The Exposed Self* (Free Press).
Llinás, Rodolfo R. (2001), *I of the Vortex — From Neurons to Self* (Cambridge, MA: MIT Press).
McFadden, Johnjoe (2001), 'Synchronous Firing and Its Influence on the Brain's Electromagnetic Field: Evidence for an Electromagnetic Field Theory of Consciousness', *Journal of Consciousness Studies*, **9** (4), pp. 23–50.
McFadden, Johnjoe (2002), 'The Conscious Electromagnetic Information (Cemi) Field Theory: The Hard Problem Made Easy?', *Journal of Consciousness Studies*, **9** (8), pp. 45–60.
Ramachandran, V.S., and Blakeslee, Sandra (1998), *Phantoms in the Brain: Human Nature and the Architecture of the Mind* (London: Fourth Estate).
Schacter, Daniel L. (1996), *Searching for Memory: The Brain, the Mind, and the Past* (New York: Basic Books).
Schacter, Daniel L. (2001), *The Seven Sins of Memory: How the Brain Forgets and Remembers* (Boston, MA: Houghton Mifflin).
Wegner, Daniel M. (1989), *The White Bear and Other Unwanted Thoughts: Suppression, Obsession, and the Psychology of Mental Control* (New York: Viking, 1989. New edition New York: Guilford Press, 1994).
Wegner, Daniel M. (2002), *The Illusion of Conscious Will* (Cambridge, MA: MIT Press).
Wright, Robert (2000), *Nonzero: The Logic of Human Destiny* (New York: Pantheon).